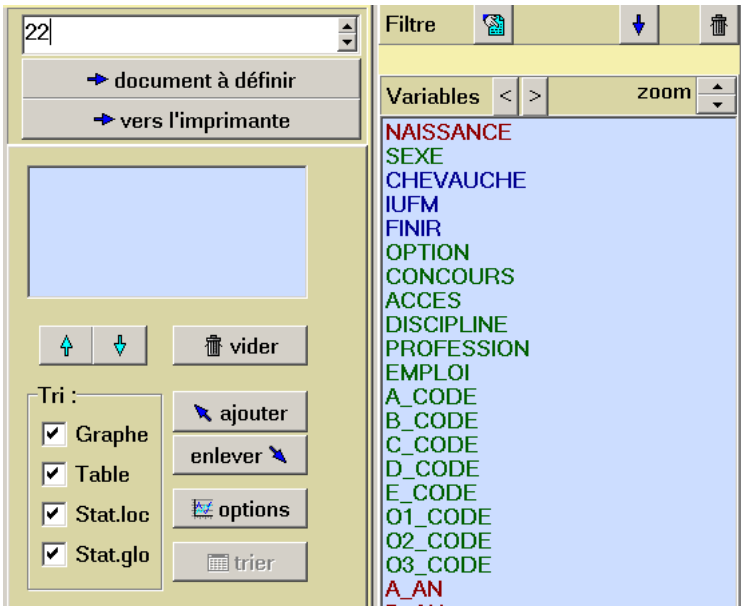
# Hector²

Manuel des traitements

# TRIS, CROISEMENTS ET TRAITEMENTS SIMPLES

## *Aspect de la page TRAITEMENTS*

On accède à la page des traitements depuis la page d'accueil, en sélectionnant l'onglet TRAITEMENTS. Elle a initialement l'aspect suivant :




On n'aperçoit ici que le coin en haut à droite de l'écran, mais il contient tout ce dont on a besoin pour le moment, c'est-à-dire les instruments des traitements simples.

## Parties communes

La partie gauche de l'écran, qu'on ne voit pas ci-dessus, est la zone d'affichage des résultats, qu'on appellera, plus simplement, l'affichage.

La partie droite de l'image comporte, de haut en bas :

- La zone d'utilisation des filtres
- La liste générale des variables
- La liste générale des collections, ici invisible, et qui ne joue aucun rôle dans les traitements simples.

Dans le secteur central, on trouve :

- Un sélecteur de taille de caractères ; il influe sur l'ensemble des proportions des tableaux et graphiques.
- Le bouton [vers document à définir], qui sert à envoyer le contenu de l'affichage vers le document de communication avec d'autres logiciels. Cette question est traitée en détail dans le chapitre « Exploitation des résultats »

On désignera comme « panneaux de commande » les surfaces regroupant les organes correspondant à chaque groupe de fonctionnalités.

Un onglet en bas du panneau de commande donne également accès au panneau « collectifs » et, dans la version professionnelle, au panneau « projection ».

Ce chapitre traite spécifiquement du panneau de commande [tris]. Les deux autres sont traités respectivement dans les chapitres « Traitements collectifs » et « Utilisation du plan de projection ».

## Description du panneau de commande [tris]

La fenêtre bleu ciel, actuellement vide, contiendra la liste (des variables) à trier, celles qui vont entrer dans le tri ou le croisement à afficher.

> La démarche de base en analyse de données consiste à considérer les cas de figure distincts (tri) et à compter combien de sujets relèvent de chacun des cas (recensement), puis à construire sur cette base des tableaux, statistiques et graphiques plus ou moins élaborés. On parle assez couramment de *tri à plat*, ou simplement *tri*, quand une seule variable est impliquée, et de *croisement* dès que plus d'une variable est impliquée, avec des distinctions entre croisement *simple* (deux variables) et *multiple* (plus de deux variables). La séparation nette entre tri à plat et croisement provient sans doute du fait que le premier relève d'une approche *descriptive* des données, tandis que le second peut introduire à une démarche *inductive* de recherche de relation entre variables. Toutefois, si on veut unifier le vocabulaire et rester rigoureux, on pourrait parler de tri sur une, deux, trois… variables.

Le panneau de commande [tris] autorise le tri simultané jusqu'à quatre variables. Pour un nombre supérieur de variables, d'autres techniques sont proposées ailleurs.

Quand des variables sont installées dans la liste à trier, on peut en modifier l'ordre (qui peut avoir des conséquences sur la disposition des résultats) en sélection une variable et en agissant sur les petites flèches bleu clair, qui partout dans Hector signifient « monter » et « descendre ».

Le bouton [vider], orné d'une poubelle, vide la liste à trier.

> Droite-cliquer (cliquer avec le bouton droit de la souris) dans la liste générale des variables a le même effet que le bouton [vider].

Le bouton [ajouter] a pour effet d'ajouter dans la liste à trier la variable actuellement sélectionnée dans la liste générale des variables (la dernière si la sélection est multiple), si elle n'y est pas déjà et si aucune autre règle ne s'y oppose.

> Double-cliquer sur le nom de la variable dans la liste générale a le même effet que le bouton [ajouter].

Le bouton [enlever] a pour effet de retirer de la liste à trier les variables qui y sont sélectionnées.

Le bouton [options] donne accès à un panneau de choix d'options qui gouvernent divers aspects de la disposition ou de l'aspect des résultats. Il sera discuté en détail ci-après.

Le bouton [trier], actif quand un tri est possible, a pour effet d'exécuter le tri des variables de la liste à trier, dans l'état où est cette liste.

> Ajouter une variable qui figure déjà dans la liste, soit par le bouton [ajouter], soit en la double-cliquant dans la liste générale, a le même effet que le bouton [trier]. Ainsi la séquence rapide la plus efficace pour obtenir un résultat est, toujours dans la liste générale, de droite-cliquer n'importe où pour vider la liste à trier, de double-cliquer sur chacune des variables souhaitées, et double-cliquer encore une fois sur la dernière pour obtenir le résultat.

Toute modification de la liste à trier vide la zone d'affichage : il faut alors à nouveau [trier].

Les effets exacts du bouton [trier] dépendent à la fois :

- Des options actuellement sélectionnées
- Du type des variables de la liste à trier
- De ce qui est coché dans le cadre [tri]

Les deux premiers points seront développés en détail plus loin.

*Les cases cochées dans le cadre [tri]*

Le cadre [tri] comporte quatre cases à cocher, qui autorisent (cochées) ou interdisent (vides) la publication de certains aspects du résultat :

- Graphe : il s'agit de la représentation graphique du tri, c'est-à-dire d'un dessin ou d'un schéma.
- Table : il s'agit de la publication des résultats au format tabulaire, c'est-à-dire sous forme de tableau.
- Stat.loc : il s'agit de la statistique locale, c'est-à-dire d'éléments statistiques qui éclairent le détail de certains éléments du tableau, mais n'autorisent pas un diagnostic général sur le tri en cours.
- Stat.glo : il s'agit de la ou des statistiques globales, c'est-à-dire des tests et mesures qui portent sur la totalité du tri.

Cette distinction entre quatre aspects résulte d'un parti pris épistémologique et didactique de l'auteur. Le modèle sous-jacent, efficace surtout dans la démarche inductive, est le suivant :

- Je croise deux variables (par exemple, parmi les élèves de 3ème d'un collège, la classe dont il font partie et leur choix d'orientation en Seconde) entre lesquelles je cherche à mettre en évidence une relation (du genre : les élèves de 3ème ZX sont proportionnellement plus nombreux que les autres à choisir l'option « Sciences Economiques »).
- Il me semble que le résultat tabulaire montre en effet des différences entre les classes, en termes de répartition selon les options. Cependant, je ne me fie pas à ma lecture du tableau,
- Et je consulte une statistique globale, en l'occurrence le Khi2, qui m'indique que la relation est effectivement significative au seuil de .01, c'est-à-dire que le seul hasard n'aurait pu fournir une répartition pareille que dans moins d'un cas sur 100.
- Dans ces conditions, j'examine avec attention la représentation graphique.
- Que je peux commenter en m'appuyant sur les indications de la statistique locale.

La philosophie fondamentale de cette approche est que, si la statistique globale ne me donne pas le feu vert à l'étape 3, j'arrête là, et je ne commente pas plus avant, parce que, d'un strict point de vue statistique, il n'y a rien à commenter.

> Une exception à cette règle d'airain sera inroduite dans le croisement d'une variable catagorielle avec une variable numérique, ou ANOVA.

Les éléments de résultats dans le panneau de commande [tris] sont organisés selon ce principe, avec parfois quelques aménagements marginaux.

Dans certains cas, les résultats tabulaires sont trop encombrants pour être publiés, et, la statistique globale étant significative, on la publiera seulement avec la représentation graphique. Dans d'autres cas, la statistique globale n'étant pas significative, on ne fournira ni la statistique locale ni la représentation graphique. Il suffit de cocher/dé-cocher les cases appropriées du cadre [tri].

## *Combinaisons de types dans les traitements simples*

Le type des variables triées ensemble détermine totalement le genre de résultats qui va être affiché. Dans le cas du tri d'une seule variable, chaque type possède son propre système de résultats.

## Tris d'une seule variable

Rappelons la définition des cinq types de variables admis par Hector (voir aussi le Chapitre « Concepts fondamentaux … ») :

- Le type calendaire (souvent noté par le symbole µ) et affiché en violet dans les listes, est une sorte de variable numérique spécialisée dans le codage des dates, et dotée à ce titre d'un format d'affichage particulier.
- Le type numérique (souvent noté par le symbole #) et affiché en rouge dans les listes, représente toutes les sortes de nombres, positifs ou négatifs, entiers ou réels.
- Le type logique (souvent noté par le symbole £) et affiché en bleu dans les listes, représente les variables binaires à valeurs Vrai/Faux, ou 1/0.
- Le type ordinal (souvent noté par le symbole §) et affiché en brun doré dans les listes, représente des variables dont les différentes valeurs sont exprimées par des étiquettes, parmi lesquelles l'ordre possède une signification (et ne peut être modifié sans altérer la sémantique de la variable).
- Le type nominal (souvent noté par le symbole $) et affiché en vert dans les listes, représente des variables dont les différentes valeurs sont exprimées par des étiquettes, parmi lesquelles l'ordre est arbitraire (et peut être modifié sans aucun dommage).
- Le type texte-libre (souvent noté par le symbole ¤) et affiché en noir dans les listes, correspond à des textes de longueur quelconque, qui ne peuvent faire l'objet d'aucune exploitation statistique, hormis dans la page des Séquences (version professionnelle seulement).

*µ : une calendaire*

Les résultats tabulaires ont l'aspect suivant :

Nb : plutôt que des copies d'écran, on utilise ici, pour un meilleur rendu, les résultats tel qu'on les récupère dans un traitement de texte par la technique du document (cf. Chapitre « Exploitation des résultats »). L'aspect est de ce fait conforme à ce que l'usager obtiendra dans ses propres travaux.

`date d'inscription`

| [classes[ | effectifs | %/Total | % cumulés |
|---|---|---|---|
| 01/01/1987 | 40 | 4,14% | 4,14% |
| 01/01/1988 | 125 | 12,93% | 17,06% |
| 01/01/1989 | 114 | 11,79% | 28,85% |
| 01/01/1990 | 106 | 10,96% | 39,81% |
| 01/01/1991 | 109 | 11,27% | 51,09% |
| 01/01/1992 | 114 | 11,79% | 62,87% |
| 01/01/1993 | 119 | 12,31% | 75,18% |
| 01/01/1994 | 127 | 13,13% | 88,31% |
| 01/01/1995 | 113 | 11,69% | 100,00% |
| Total | 967 | 100.00% | |

La première colonne décrit les classes utilisées dans ce tableau qui résume le tri ; les valeurs individuelles sont trop nombreuses pour être représentées toutes en détail : Hector utilise donc des classes. Comme il s'agit d'une variable calendaire, les classes sont spécialement calculées pour s'aligner ici avec le début d'une année, mais avec une moindre étendue des classes représentées, ce pourrait aussi bien être le mois ou la semaine. Ces classes débutent avec la valeur affichée dans

la ligne, incluse, et s'arrêtent avec la valeur de la ligne suivante, exclue, ce qu'exprime le titre [classes[. Ainsi la classe notée 01/01/1988 doit s'entendre comme [01/01/1988, 01/01/1989[ : tout ce qui est entre le premier janvier 1988 et le premier janvier 1989, à l'exclusion de ce dernier, autrement dit tous les jours de 1988.

La seconde colonne affiche les effectifs, c'est-à-dire le nombre de sujets relevant de chaque classe. Ces deux premières colonnes constituent le format tabulaire brut dans ce cas de figure.

En effet, les deux colonnes suivantes sont ici l'expression de la sélection des *statistiques locales*. La troisième colonne présente les pourcentages par classe, autrement dit les effectifs rapportés à l'effectif total (et multipliés par 100). La quatrième colonne présente les pourcentages cumulés : chaque cas contient le pourcentage de sujets obtenu en additionnant les sujets relevant de cette classe et de toutes celles qui précèdent. On lit ainsi qu'en prenant toutes les classes depuis le début jusqu'à [ 01/01/1991 [, on obtient 51,09% des sujets, soit à peu près la moitié.

Une remarque très importante doit être faite au sujet de l'effectif total : il est ici de 967, ce qui est en effet le total de sujets de ce corpus. Dans d'autres tableaux et résultats, l'effectif total pourra être inférieur à ce nombre, si le tri implique des variables qui ne sont pas définies pour tous les sujets (pour lesquelles certains sujets ont une non-valeur). Dans ce cas, les sujets possédant une non-valeur pour au moins une des variables impliquées dans le tri ne peuvent apparaître dans le résultat et n'y sont pas totalisés, même pas comme non-réponses. C'est à l'usager de connaître son corpus et d'indiquer, si cela lui paraît nécessaire, que x sujets sont exclus du tableau pour ce genre de raison.

Voici les statistiques globales du même exemple :

```
Classe modale : [01/01/1994,01/01/1995[ (n=127)
Médiane entre 19/11/1991 & 20/11/1991
M : 01/11/1991 M-s : 31/05/1989 M+s : 02/04/1994
```

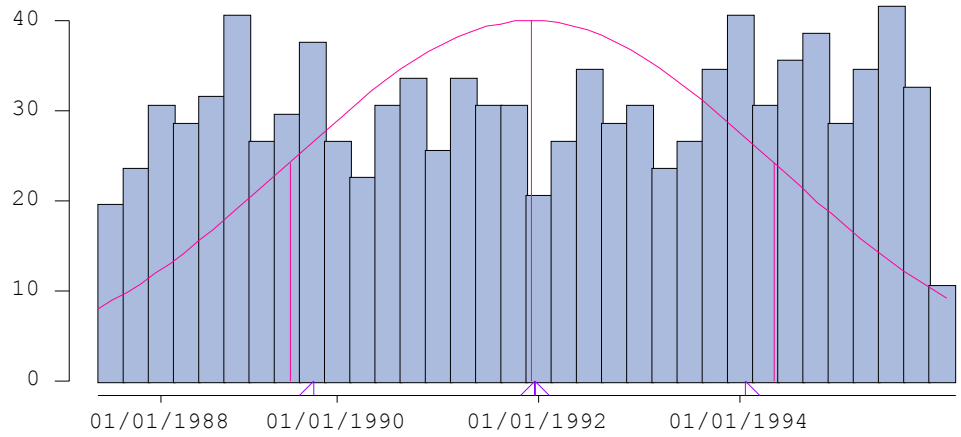
La classe modale, celle qui rassemble le plus grand nombre de sujets parmi les classes, est prises parmi les classes au sens du type de regroupement opéré dans le tableau (on verra que cela est paramétrable). Elle est ici exprimée complètement avec ses deux bornes, l'incluse à gauche et l'exclue à droite, et l'effectif correspondant.

La médiane n'est pas liée à une valeur, mais représente une coupure entre deux valeurs existantes, coupure réalisée de manière à ce que les deux parties résultantes de cette coupure soient aussi proches que possible de l'équilibre 50/50 : c'est la frontière, le coup de rasoir qui sépare la distribution en deux masses approximativement égales. Ici elle passe entre deux jours, parce que les variables calendaires utilisent le jour comme unité insécable.

> Certains auteurs d'ouvrages statistiques considèrent qu'il y a lieu d'interpoler entre deux valeurs existantes, pour découvrir une valeur virtuelle qui couperait exactement la distribution en deux. L'auteur du présent logiciel n'adhère pas à cette manière de voir, et s'en tient à la théorie de la coupure entre deux valeurs. Cette question a à voir avec les considérations de continuité ou de discontinuité des variables numériques, un casse-tête en mathématiques, mais qui en informatique n'a pas lieu d'être parce que toutes les représentations informatiques de nombres sont, *in fine*, digitales et donc discontinues.

Enfin la dernière ligne fournit la moyenne, ainsi que la moyenne plus l'écart-type et moins l'écart-type. S'agissant de données calendaires, toutes ces valeurs sont ramenées à l'unité de jour la plus proche.

La représentation graphique correspondant au tri de la variable calendaire est la suivante :




Il s'agit, comme on le verra ci-après, d'une simple variante de la représentation graphique des numériques : un histogramme. L'échelle verticale compte le nombre de sujets représentés par chaque colonne. Les marques pourpres notent la moyenne, les deux écarts-types et l'allure qu'aurait la distribution, avec ces paramètres, si elle était normale. De même, les marques violettes repèrent, au centre, la médiane, et, de part et d'autre, les deux autres frontières de quartiles (coupures qui découpent l'effectif en quatre masses approximativement égales).

Parmi les options auxquelles on accède par le bouton [options], celles qui concernent le tri d'une calendaire sont les suivantes :

- La valeur [tri d'une variable] dans le cadre [nombre maximal de lignes], lui-même dans le cadre [contraintes de dimension]. Ce paramètre, en jouant sur le nombre maximal de lignes du tableau, influe directement sur le niveau de regroupement dans les classes de valeur.
- La valeur [histogramme], dans le cadre [nombre maximal de colonnes], dans le même secteur [contraintes de dimensions]. Ce paramètre limite le nombre de colonnes distinctes dans l'histogramme et influe donc directement sur le niveau de regroupement des valeurs dans ce graphique.
- Le sélecteur [couleurs des graphiques], dans le cadre [Paramètres d'aspect]. Les possibilités distinctes sont assez nombreuses : l'usager est invité à essayer de découvrir un style qui lui convienne.

*# : une numérique*

D_NOTE

| [classes[ | effectifs | %/Total | % cumulés |
|---|---|---|---|
| 0.00 | 6 | 0,95% | 0,95% |
| 2.00 | 8 | 1,27% | 2,22% |
| 4.00 | 16 | 2,54% | 4,75% |
| 6.00 | 18 | 2,85% | 7,61% |
| 8.00 | 33 | 5,23% | 12,84% |
| 10.00 | 98 | 15,53% | 28,37% |
| 12.00 | 98 | 15,53% | 43,90% |
| 14.00 | 125 | 19,81% | 63,71% |
| 16.00 | 129 | 20,44% | 84,15% |
| 18.00 | 81 | 12,84% | 96,99% |
| 20.00 | 19 | 3,01% | 100,00% |
| Total | 631 | 100.00% | |

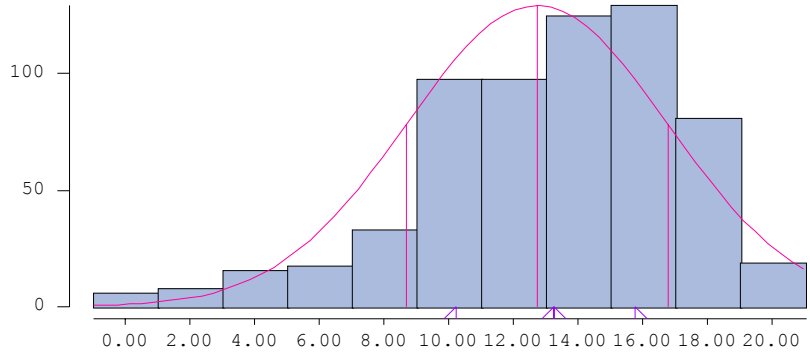
L'aspect du tableau, avec les pourcentages simples et cumulés comme statistique locale, ne diffère pas beaucoup de la variable calendaire. Tout au plus peut-on observer que les valeurs de la

variable sont affichées avec deux zéros après la virgule, parce que cette variable est définie avec deux décimales.

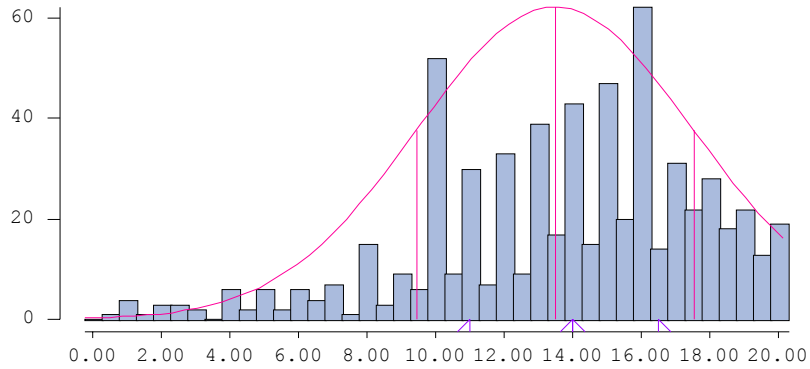
```
Classe modale : [16.00,18.00[ (n=129)
Médiane entre 14.00 & 14.25
Moyenne 13.709, écart-type 4.040
Coupure contrastée après 12.500
H(normalité) rejetée à .0000 ; H(symétrie) rejetée à .001
```

De la même manière, la position de la médiane entre les valeurs 14.00 et 14.25 nous indique que cette variable va par quarts de point ; les classes du tableau et de l'histogramme, qui vont de deux points en deux points, contiennent donc en fait huit valeurs d'origine.

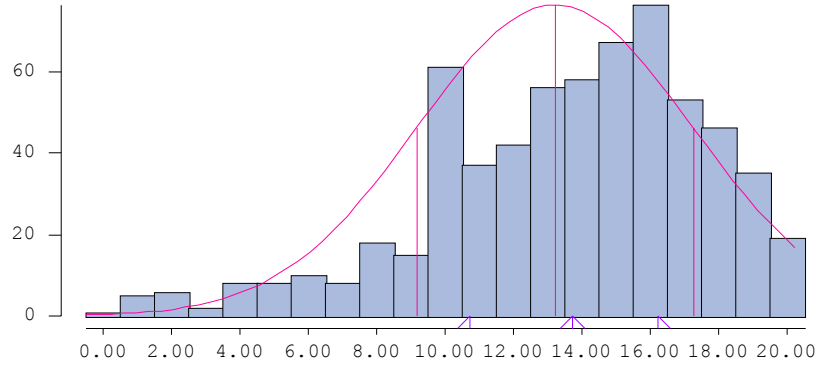
Les deux dernières lignes sont dépendantes de choix opérés dans le panneau d'Options. La coupure contrastée est celle qui découpe la distribution en deux sous-ensemble tels qu'entre eux le contraste, mesuré par le |t| de Student, soit maximum. A dernière ligne pose le test de normalité et de symétrie. Pour plus de détails, se reporter à l'Annexe Statistique.




Les options jouent ici le même rôle qu'avec la variable calendaire. En poursuivant la réflexion sur la définition de l'intervalle minimum dans la variable étudiée, on pourrait souhaiter obtenir une définition plus fine de l'histogramme. Le suivant est obtenu avec une valeur de 50 pour le paramètre « nombre maximal de colonnes dans un histogramme ». La précision résultante de l'histogramme est en demi-points. Elle n'est pas très intéressante à cause de la dispersion due au fait que toutes les décimales ne sont pas utilisées avec la même probabilité.



Une autre tentative, avec un nombre de colonnes maximal de 30, donne ce compromis assez satisfaisant, avec une résolution d'un point entier pour l'histogramme :




On le voit, le meilleur résultat dépend du genre de données dont on dispose, et sa recherche peut donner lieu à quelques essais et erreurs. Pour autant, on a choisi de fixer ce genre d'options de manière externe, transversale et non pas liée à chaque tableau, car il convient de traiter de la même façon tous les tableaux concernant des variables du même genre. L'usager a donc intérêt à procéder à un réglage global qui convienne à une série de variables, avant d'exporter ses résultats.

### *£ : une logique*

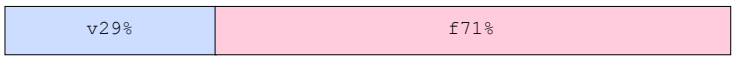
IUFM

| | effectifs | %/Total |
|---|---|---|
| Iufm- | 685 | 70.84% |
| Iufm+ | 282 | 29.16% |
| Total | 967 | 100.00% |

Le tableau ressemble aux précédents. La statistique globale est ici réduite à sa plus simple expression : le pourcentage. Le pourcentage cumulé ne présente en effet aucun intérêt. On notera la formulation des légendes des lignes du tableau : des étiquettes sont fabriquées avec le nom de la variable et un suffixe – pour faux, + pour vrai. Or une étiquette ne saurait dépasser 10 caractères. En enlevant le suffixe, il n'en reste que 9, alors qu'un nom de variable peut comporter 35 caractères. Hector utilise dans ce cas son algorithme d'abréviation, qui par exemple pour {Inscription annuelle régulière}, valeur {Vrai}, donnerait l'étiquette {InsAnnRég+}. C'est un élément dont il faut tenir compte quand on choisit les noms des variables logiques.

Intervalle de confiance à .05 : [27.86% 30.46%]

La statistique globale est ici l'intervalle de confiance, qui s'interprète ainsi : le taux de vrais 29,16 %, constaté sur cet échantillon, peut être considéré comme représentatif, avec moins de 5% d'erreur, d'une « vraie valeur » comprise entre 27,86% et 30,46%. Cette notion s'apparente à celle de fourchette dans les sondages. Le graphique est une simple barre découpée selon les proportions des taux de vrais et de faux.




La seule option qui agisse dans le cas d'une seule variable logique est le sélecteur [couleurs des graphiques], dans le cadre [Paramètres d'aspect].

### *§ : une ordinale*

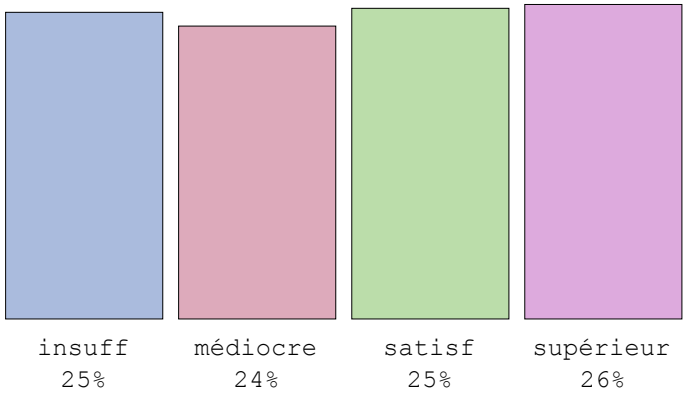
classe de résultats

| | effectifs | %/Total | % cumulés |
|---|---|---|---|
| insuff | 201 | 25.13% | 25.13% |
| médiocre | 191 | 23.88% | 49.00% |
| satisf | 203 | 25.38% | 74.38% |
| supérieur | 205 | 25.63% | 100.00% |
| Total | 800 | 100.00% | |

Le tableau ressemble à celui des numériques, sauf qu'il ne peut être question de classes de valeurs : les valeurs d'une ordinale sont représentées par des textes arbitraires que rien n'autorise à regrouper. Toutefois, on s'autorise ici les pourcentages cumulés : c'est qu'à la différence d'une variable nominale, il existe un ordre significatif entre les valeurs. {supérieur} est, d'une certaine manière, après {satisf}, qui est lui-même après {médiocre}, et on a le droit d'en déduire que {supérieur} est après {médiocre}. En revanche, on ne peut pas dire que la différence entre {médiocre} et {satisf} soit en quoique ce soit comparable à la différence entre {satisf} et {supérieur}, sinon on serait devant une échelle d'intervalles, proche d'une numérique. Ce caractère ordinal est tout ce qui distingue une variable ordinale d'une variable nominale. Il permet des tests statistiques plus riches qu'une variable nominale, ne serait-ce que parce qu'il admet la notion de rang : des 205 sujets notés {supérieur}, on pourrait dire qu'ils sont premiers *ex aequo*, parce qu'ils sont meilleurs que tous les autres ; plus précisément, ils occupent ensemble les rangs allant du 1 au 205 ; il est donc plus juste de dire qu'ils sont classés ensemble au 103ème rang (103 = (205+1)/2).

```
Valeur modale : supérieur (n=205)
Médiane entre médiocre & satisf
```

C'est aussi ce caractère ordinal qui autorise la définition d'une médiane : si l'on veut couper cette distribution aussi près que possible de l'équilibre 50/50, c'est entre {médiocre} et {satisf} qu'il faut le faire : cela donne 49/51 très exactement, comme on peut le lire dans la colonne des pourcentages cumulés.




La représentation graphique, sans chercher à rappeler un histogramme, exprime quand même de gauche à droite la progression ordinale des valeurs, ce qu'on se gardera de faire avec une variable nominale. Si le nombre de valeurs distinctes de la variable ordinale excède huit, et pour éviter des graphiques trop larges, le graphique utilise le modèle du « graphe en barres », l'une des options pour les variables nominales.

La seule option active ici est le sélecteur [couleurs des graphiques], dans le cadre [Paramètres d'aspect].

*$ : une nominale*

OPTION

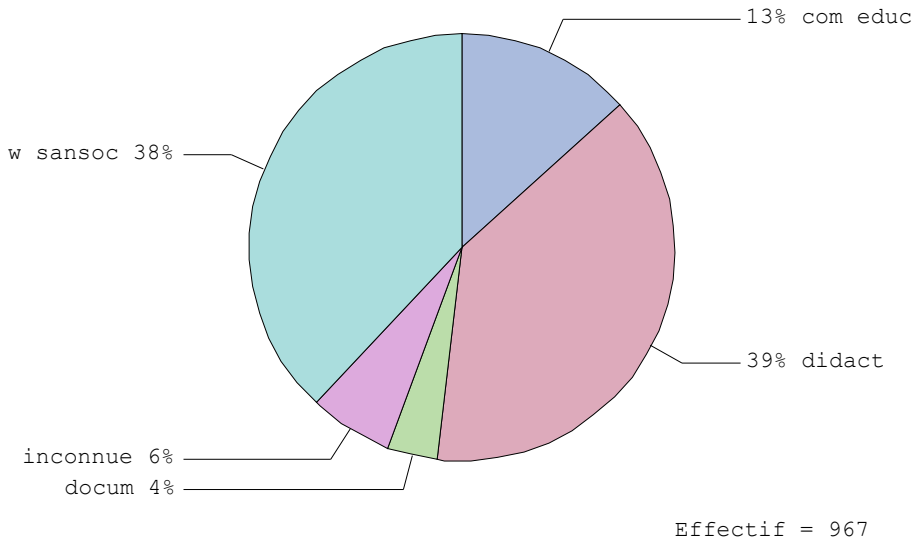
| | effectifs | %/Total |
|---|---|---|
| com educ | 129 | 13,34% |
| didact | 373 | 38,57% |
| docum | 36 | 3,72% |
| inconnue | 62 | 6,41% |
| w sansoc | 367 | 37,95% |
| Total | 967 | 100.00% |

Le résultat tabulaire et la statistique locale sont similaires à la variable ordinale, sauf qu'il n'y a pas de colonne de cumul des pourcentages, qui n'aurait aucune signification du fait de l'absence de toute relation d'ordre. La statistique globale, purement descriptive, repose sur une comparaison entre l'entropie de la distribution, au sens de la théorie de l'information, et l'entropie idéale, qui

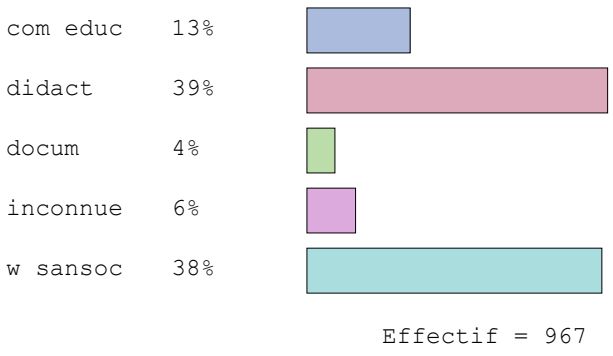
serait mesurée dans le cas d'équiprobabilité des valeurs. Cette mesure est apparentée à celle du rendement d'un canal de transmission de signaux. Elle n'est pas dépourvue d'intérêt pour décider de regrouper ou non des catégories dans une variable nominale (via une dérivation). Une variable à faible efficacité entropique se manifeste par une énorme catégorie majoritaire, accompagnée d'une poussière de petites catégories. Telle quelle, une variable de ce genre ne permettra pas de mettre en évidence des relations statistiques.

```
Efficacité entropique : 80,9%
```

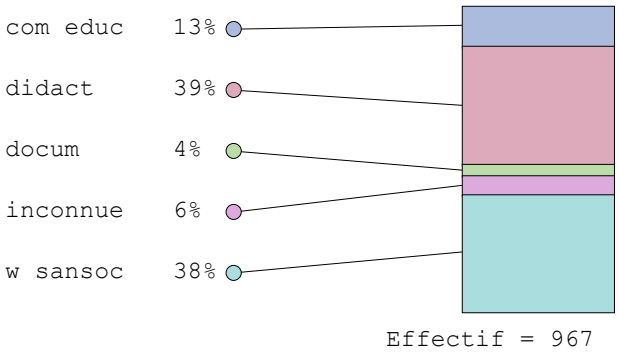
Le genre de graphique par défaut (initialement sélectionné) pour une nominale est le graphique en secteurs (alias camembert). Chaque secteur possède un angle (et donc une surface) proportionnel à la part d'effectif concerné par la valeur affichée.




Cette représentation, par son caractère cyclique, vise à manifester l'absence de relation d'ordre dans une variable nominale. Dans certains cas, cependant, elle n'est pas satisfaisante, notamment dans le cas où les positions trop nombreuses engendrent un fouillis de secteurs et d'étiquettes illisibles (voire, sur certains systèmes, un problème de ressources graphiques). On dispose pour ces cas de deux autre possibilités :

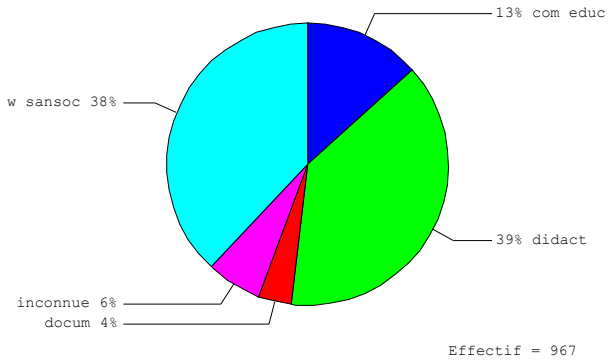


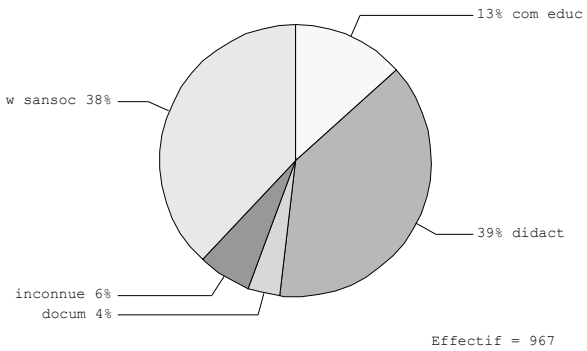

Celle-ci est le diagramme en barres. C'est la longueur des barres qui est proportionnelle aux effectifs des catégories. Ce modèle est également utilisé pour les variables ordinales quand leur nombre de catégories dépasse 8.

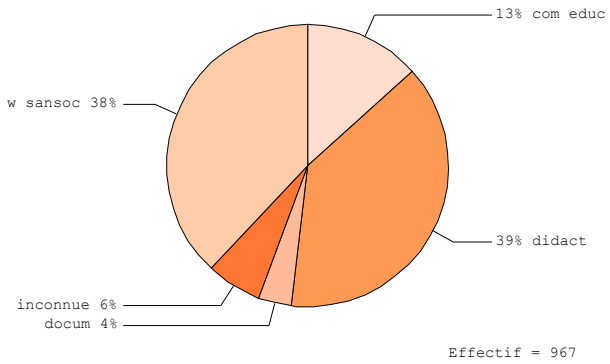



Ci-dessus, le diagramme en couches. Cette fois, c'est l'épaisseur de chaque couche qui est proportionnelle à l'effectif. Cette présentation est bien adaptée aux variables nominales qui ont beaucoup de positions différentes.

Parmi les options, outre le paramètre [graphe de nominale] (barres, couches, secteurs), qui détermine le modèle de rendu graphique de la variable nominale, le sélecteur [couleurs des graphiques], dans le cadre [Paramètres d'aspect], joue un rôle dans cette circonstance :






Ci-dessus, quelques échantillons :

- avec des couleurs vives
- avec un système de grisés
- dans l'un des douze systèmes de camaïeu disponibles.

Un système de hachures est également disponible, mais il fournit des résultats assez laids, et doit être réservé aux cas on on dispose vraiment d'une trop mauvaise imprimante.

## Tris de deux variables

On les appelle aussi croisements. Sachant qu'il existe cinq types de variables statistiquement utiles, on pourrait s'attendre à 25 modèles de tris de deux variables. En fait, par le jeu de différentes règles et d'assimilations, le nombre de cas de figures est nettement plus réduit.

Si l'on représente chaque type par son symbole, on obtient la table d'assimilation ci-dessous :

| type | | µ | # | £ | § | $ |
|---|---|---|---|---|---|---|
| Calendaire | µ | ## | ## | *$#* | *$#* | *$#* |
| Numérique | # | ## | ## | *$#* | *$#* | *$#* |
| Logique | £ | $# | $# | *££* | *§§* | $$ |
| Ordinale | § | $# | $# | §§ | §§ | $$ |
| nominale | $ | $# | $# | $$ | $$ | $$ |

Voici les deux règles qui permettent de construire ce tableau :

- Une variable numérique (ou calendaire) ne peut apparaître avant une variable catégorielle (logique, ordinale, nominale), parce que ça n'est pas intéressant du point de vue des modèles de représentation. Si c'est le cas, on inverse l'ordre des variables. De telles inversions sont signalées par de l'italique dans le tableau ci-dessus.
- Quand deux variables catégorielle de « vertu » inégale sont croisées, le modèle dépend de la variable dont la vertu est la plus faible : ainsi une ordinale croisée avec une nominale se comporte elle-même comme une nominale : §$->$$.

On se retrouve donc, au lieu de 25, en présence de 5 cas de figure :

- ## croisement de numériques ou assimilées
- *££* croisement de logiques
- §§ croisement d'ordinales
- $$ croisement de nominales
- $# croisement mixte, souvent désigné comme ANOVA (Analysis of variance)

*## croisement de numériques*

C_NOTE × E_NOTE

| | 0.00 | 5.00 | 10.00 | 15.00 | 20.00 | S/LIGNE : |
|---|---|---|---|---|---|---|
| 0.00 | 7 | 3 | 21 | 7 | | 38 |
| 5.00 | 5 | 12 | 73 | 46 | | 136 |
| 10.00 | 7 | 7 | 153 | 77 | | 244 |
| 15.00 | 1 | 4 | 22 | 25 | 1 | 53 |
| 20.00 | | 1 | 2 | | | 3 |
| S/COLONNE: | 20 | 27 | 271 | 155 | 1 | 474 |

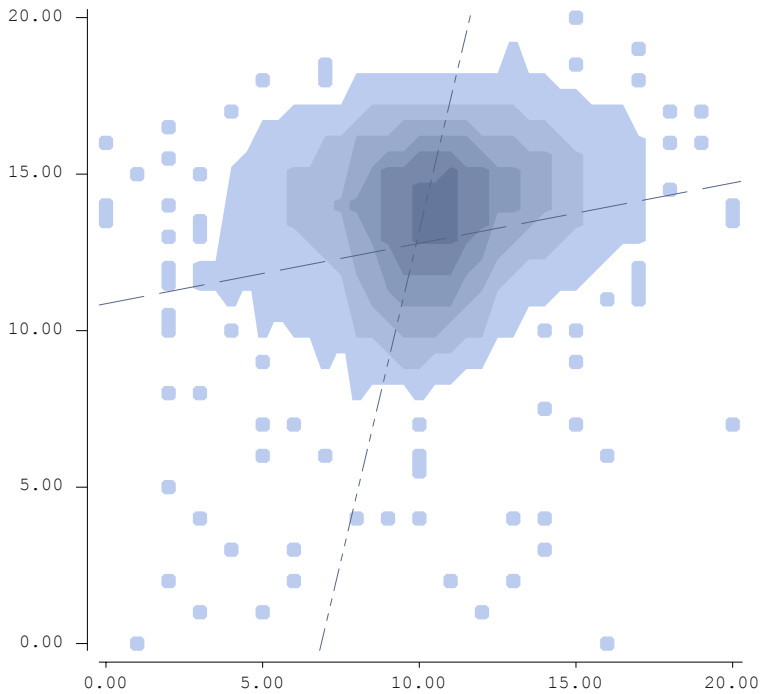
Le format tabulaire est commun à tous les croisements de deux variables, hormis l'ANOVA ($#). Chaque case contient l'effectif concerné simultanément par la valeur de la ligne et par celle de la colonne auxquelles elle appartient.

La statistique locale est également commune à tous les croisements de deux variables, et propose, selon l'option choisie, les pourcentages ligne ou colonne ou le signe des associations locales (voir ci-dessous). On n'a fait figurer aucune statistique locale dans cet exemple, parce que ces éléments sont de peu d'intérêt pour un croisement de numériques (exceptionnellement, ils peuvent en avoir, d'où leur disponibilité).

Le tableau ci-dessus n'est compact que parce qu'on a utilisé, dans les options, des contraintes de nombre de lignes et de colonnes drastiques, de manière à réduire le nombre de classes, pour être sûr de tenir dans la page. En fait, sauf exception, on n'utilise guère le format tabulaire des croisements de numériques, qui génère volontiers des tableaux immenses et illisibles.

La statistique du coefficient de corrélation r de Bravais-Pearson est le choix standard dans le cas d'un croisement de numériques. La valeur obtenue ici, 0.214, est significative au seuil de .05 : compte tenu de l'importance de l'effectif, c'est un résultat assez modeste.

r (Bravais-Pearson) = 0.214 , s. à .05




La représentation graphique rappelle horizontalement la variable en lignes et verticalement la variable en colonnes. Les zones de couleur cherchent à constituer un nuage de densité : plus les zones sont sombres, plus la population est dense dans cette région.

Les deux droites sont les droites de régression d'y en x et d' x en y . Elles représentent les relations entre deux variables, dans les deux hypothèses où l'une serait entièrement dépendante de l'autre à une certaine quantité de bruit (erreur) près. Le fait qu'elles forment un ciseau assez

ouvert est à relier à la valeur peu élevée du coefficient de corrélation. Avec un coefficient proche de 1, les deux droites seraient presque confondues sur la diagonale principale (la bissectrice de l'angle origine) . Avec un coefficient proche de 0, elles seraient quasi perpendiculaires, l'une horizontale, l'autre verticale. Avec un coefficient proche de –1, elles seraient à nouveau presque confondues, mais sur la contre-diagonale (perpendiculaire à la principale).

D'assez nombreuses options gouvernent l'aspect des résultats d'un croisement de numériques :

- Dans le cadre Options Statistiques, le cadre [stat. Locale des tableaux croisés] offre le choix entre pourcentages en lignes, pourcentages en colonnes et signe des associations locales, si une stat. Locale est demandée.
- Le cadre [stat. Globale pour les croisements] propose, à la ligne [de numériques], les choix suivants (pour le lecteur qui ne connaîtrait pas la signification et l'usage de ces mesures, il est possible de se reporter à l'annexe « Abrégé statistique ») :
  - rBP : le coefficient de corrélation r de Bravais-Pearson, proposé par défaut.
  - reg : les coefficients des équations des droites de régression
  - rhô : le coefficient de corrélation par rangs d Spearman
  - |t|a : le |t| de Student sur échantillons appareillés
- Dans le cadre [paramètres d'aspect], le sélecteur [couleur des graphiques]
- Dans le cadre [nuage de densités],
  - La granularité fixe le nombre maximum de cellules élémentaires de couleur en x et en y (le « grain » du schéma)
  - Le rayon de lissage indique sur combien de voisines à la ronde une cellule donnée étend son influence
  - Le paramètre de seuillage indique combien de zones distinctes de densité il faut représenter
  - L'option [accentuer le contraste] est utile dans certains nuages aux zones tendanciellement indifférenciées.
- Les contraintes de dimensions en lignes et colonnes pour l'éventuel tableau

Le choix des options convenables est à la fois affaire de préférences esthétiques et d'adaptation au données effectivement analysées.

### *£.£. croisement de logiques*

IUFM × FINIR

| %LIGNE | Finir- | | Finir+ | | S/LIGNE : | |
|---|---|---|---|---|---|---|
| Iufm- | 291 | 42% | 394 | 58% | 685 | 100% |
| Iufm+ | 25 | 9% | 257 | 91% | 282 | 100% |
| S/COLONNE: | 316 | 33% | 651 | 67% | 967 | 100% |

On a ici choisi, comme statistique locale, des pourcentages en lignes, qui sont appropriés à commenter l'hypothèse selon laquelle la variable en lignes serait la variable indépendante et la variable en colonnes, la variable dépendante, autrement dit que {FINIR} dépend de {IUFM}. Dans ce contexte, on lit directement que 58% des Iufm- sont aussi Finir+, alors que c'est le cas de 91% des Iufm+.

> Traduction pour ceux qui aiment comprendre de quoi il s'agit. Ce sont des données sur des étudiants d'une licence de Sciences Humaines. Ce que dit le tableau, c'est qu'un très grande majorité de ceux qui déclarent avoir l'intention de se présenter au recrutement de l'Institut Universitaire de Formation des Maîtres déclarent également qu'ils comptent finir leur licence cette année-là.

Cette disposition, indépendante en lignes, dépendante en colonnes, pourcentages en lignes, est assez habituelle pour faire figure de convention, au point qu'il est préférable de signaler les cas où, à cause du nombre de lignes ou de colonnes, on est obligé de permuter les deux variables et d'utiliser les pourcentages en colonnes.

La relation observée dans le tableau est confirmée par la statistique globale, ici le Khi2, proposé par défaut.

```
Khi2 = 102.61 pour 1 d.d.l. , s. à .01
```

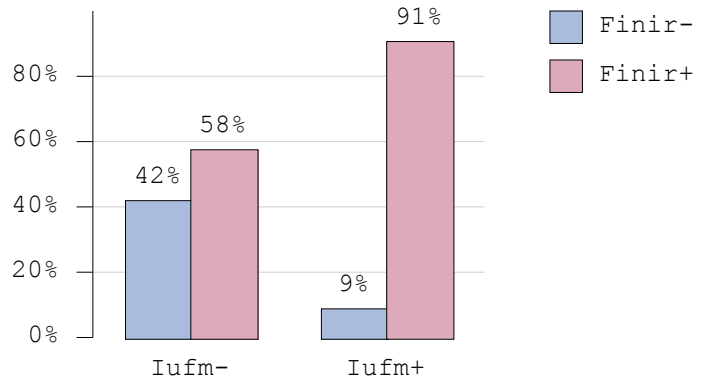
Dans le cas d'un croisement de logiques, d'autres statistiques globales sont possibles :

- rBP : le coefficient de corrélation typique des numériques
- rhô : le coefficient de corrélation par rangs, pour les numériques et les ordinales
- Cn, le coefficient normé de contingence, issu du Khi2 (typique des nominales)
- => : l'implication, coefficient défini exclusivement sur les logiques
- gGK, le coefficient de co-ordonnancement gamma de Goodman-Kruskal (ordinales)
- les statistiques de sensibilité, spécificité, prédictibilité, très utiles dans l'analyse des tests

On trouvera là aussi des définitions détaillées dans l'annexe statistique. La profusion de statistiques globales possibles avec des variables logiques illustre bien à quel point ces variables bénéficient des vertus de tous les autres types (sauf calendaire) : la variable logique admet la corrélation, si on l'interprète comme une numérique à valeurs 0/1 ; n'ayant que deux positions, elle est nécessairement ordinale ; ayant des positions repérées par des étiquettes, elle peut fonctionner comme une nominale. Qui plus est, elle possède un coefficient en propre : l'implication, ainsi que l'analyse de prédictivité.

La représentation graphique qui serait normale ici serait l'analyse factorielle simple (cf. croisement de nominales), mais celle-ci ne peut avoir pas plus de degrés de liberté (de dimensions) que celle de ses variables d'origine qui en a le moins. Le degré de liberté pour une variable de ce type se définissant comme le nombre de positions distinctes, moins 1, il serait ici de 1 : le plan de l'analyse factorielle n'aurait qu'une seule dimension, et se réduirait donc à une droite, ce qui du point de vue graphique n'est pas très expressif.

Dans un tel cas, Hector sélectionne automatiquement le mode graphique alternatif à l'AFC : le schéma en barres de pourcentages : pour chaque catégorie de la variable en lignes, on trouve autant de barres verticales que la variable en colonnes possède de catégories, avec le pourcentage de chacune.



### *§§ croisement d'ordinales*

classe d'âge × classe de résultats

| %LIGNE | insuff | médiocre | satisf | supérieur | S/LIGNE : |
|---|---|---|---|---|---|
| vétérans | 42 21% | 46 23% | 60 29% | 56 27% | 204 100% |
| mûrs | 63 30% | 55 26% | 48 23% | 46 22% | 212 100% |
| jeunes | 58 34% | 41 24% | 40 23% | 32 19% | 171 100% |
| benjamins | 38 18% | 49 23% | 55 26% | 71 33% | 213 100% |
| S/COLONNE: | 201 25% | 191 24% | 203 25% | 205 26% | 800 100% |

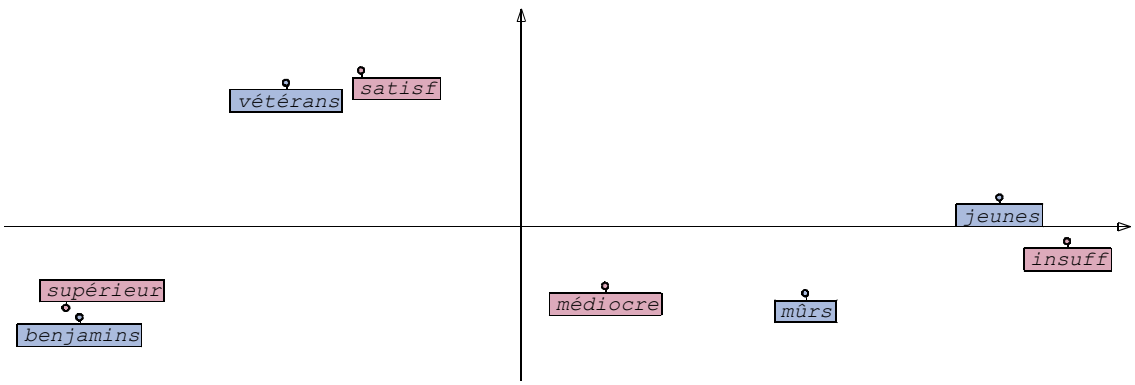
Le format tabulaire est le même que précédemment. La statistique locale est ici représentées par les pourcentages en lignes. On aurait pu utiliser aussi le signe des associations locales, comme ci-après pour les nominales.

La statistique globale fournit ici un exemple bien intéressant de la nécessité, parfois, de nuancer les interprétations.

```
Khi2 = 25,69 pour 9 d.d.l. , s. à .01
rhô (Spearman) = 0,028 , n.s.
```

Bien que la statistique par défaut pour le croisement d'ordinales soit le rhô de Spearman, on a aussi utilisé le Khi2, qui est la statistique par défaut pour les nominales. L'examen de la représentation graphique ci-dessous, qui est l'Analyse Factorielle de Correspondances simple, permet de comprendre pourquoi.

```
(classe d'âge) x (classe de résultats)
```




```
93,12% de l'inertie sur l'axe 1 horizontal
6,00% de l'inertie sur l'axe 2 vertical
```

L'utilisation du rhô de Spearman, qui est un coefficient de corrélation par rangs, est légitime à cause du caractère ordinal des données. Il fonctionne ici comme un coefficient de co-ordination (on aurait pu utiliser aussi le gamma de Goodman-Kruskal). Le Khi2, lui, ne tient aucun compte de l'ordre des valeurs, il ne cherche que des associations.

L'axe principal de l'Analyse Factorielle, qui porte 93,12% de l'inertie du nuage, porte les projections des valeurs de la variable {classe de résultats} dans l'ordre. Cependant, les valeurs de la variable {classe d'âge} ne sont pas classées correctement (de leur point de vue) : il faudrait pour que ce soit le cas permuter {jeunes} et {vétérans}. Qu'en conclure ? Qu'il y a bien une forte relation entre {classe d'âge} et {classe de résultats}, ce que dit le Khi2 significatif à .01, mais que cette relation n'est pas un co-ordonnancement, ce que dit le rhô non significatif : les deux ordinalités ne sont pas apparentées.

Les options concernées par le croisement d'ordinales sont :

- Le choix de la statistique locale des tableaux croisés

- Le choix de la statistique globale : rhô (par défaut), Khi2, CnC (coefficient normé de contingence, issu du Khi2 avec neutralisation de l'effet de la forme du tableau et des effectifs), gGK (gamma de Goodman-Kruskal, coefficient de co-ordonnancement par paires de valeurs)
- La couleur des graphiques

En revanche le nombre de lignes et de colonnes ne sont pas contrôlées (sauf à dépasser 100), parce le logiciel ne peut décider de regroupements, contrairement aux cas des variables numériques.

*$$ croisement de nominales*

ACCES × OPTION

| S.ASS.LOC. | com educ | didact | docum | inconnue | w sansoc | S/LIGNE : |
|---|---|---|---|---|---|---|
| aut.lic | 4 | 27 +++ | 1 | | 3 --- | 35 |
| B+2 tk | 23 +++ | 47 + | 5 | --- | 24 --- | 99 |
| B+3 ss | 6 | 7 --- | 1 | - | 31 +++ | 45 |
| DEUG | 65 ++ | 210 +++ | 22 ++ | --- | 108 --- | 405 |
| divers | 3 ++ | 2 | 1 | | 2 | 8 |
| expe.pro | 23 | 68 ++ | 5 | --- | 48 | 144 |
| inconnu | 5 --- | 12 --- | 1 --- | 62 +++ | 151 +++ | 231 |
| S/COLONNE: | 129 | 373 | 36 | 62 | 367 | 967 |

La différence entre les ordinales et les nominales est que cette dernière ne connaissent aucun ordre : on pourrait permuter les lignes, aussi bien que les colonnes, dans le tableau ci-dessus, sans altérer en rien sa signification.

La statistique globale utilisée ici est le Khi2, la seule qui s'applique à ce cas (avec le CnC qui en est dérivé). La statistique locale utilisée ci-dessus n'est pas non plus sans relation avec le Khi2 : il s'agit du signe des associations locales, qui s'appuie sur la contribution de chaque case du tableau au total du Khi2. Celle ci peut être :

- Très forte, de manière significative à .01 : signes +++
- Forte, de manière significative à .05 : signes ++
- Assez forte, de manière significative à .10 : signe +
- Normale : pas de signe
- Assez faible, de manière significative à .10 : signe -
- Faible, de manière significative à .05 : signes --
- Très faible, de manière significative à .01 : signes ---

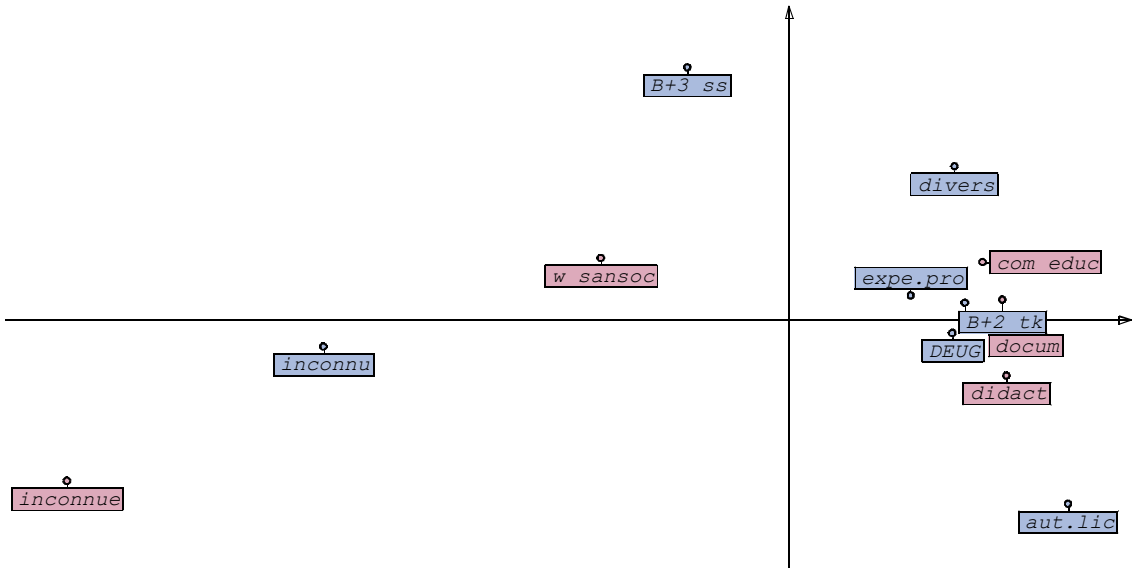
On ne doit donc pas s'étonner de voir --- dans une case vide (d'effectif nul) : un effectif nul peut être quelque chose de beaucoup plus faible que ce qui est attendu dans la case, sous l'hypothèse d'indépendance des variables, d'après les totaux ligne et colonne.

L'attention de l'usager doit être attirée sur le fait que le signe des associations locales, en tant que statistique locale, permet d'enrichir le commentaire, à condition que la statistique globale soit significative. Les petits signes à eux seuls ne permettraient pas de conclure.

Khi2 = 388.74 pour 24 d.d.l. avec 8 corrections de Yates, s. à .0000

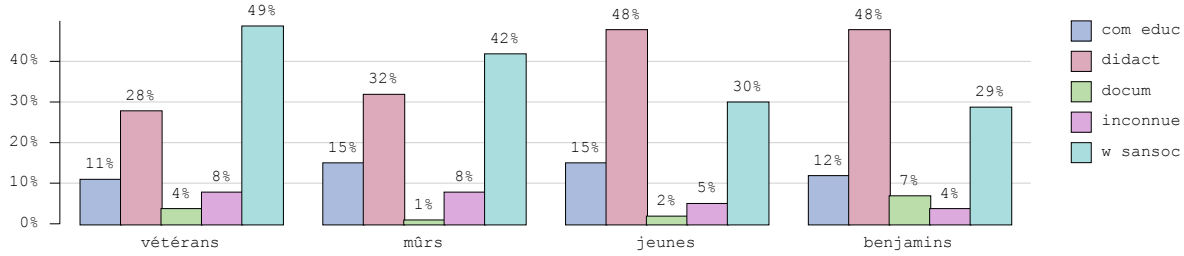
Dans cet exemple, la statistique globale du Khi2, extrêmement significative, permet tout à fait de poursuivre le raisonnement : le lecteur pourra vérifier la proximité entre étiquettes des valeurs lignes et étiquettes des valeurs colonnes, dans l'Analyse Factorielle de Correspondance qui est ici la représentation graphique, va de pair avec un ou plusieurs signes + dans le tableau : il s'agit en fait de deux manière différentes de manifester la même chose : l'association de cette ligne à cette colonne, c'est-à-dire la propension de la case de leur intersection à contenir plus de sujets que le hasard ne le laisserait prévoir.

(ACCES) x (OPTION)




91,43% de l'inertie sur l'axe 1 horizontal
6,05% de l'inertie sur l'axe 2 vertical

Une représentation alternative est le schéma en barres et %, assez vite encombrant quand les catégories sont nombreuses, mais non dépourvu d'intérêt, surtout quand l'une des variables est ordinale, et que la succession de gauche à droite a donc une signification :




Outre ce choix de représentation graphique, les options concernées par le croisement de nominales sont :

- Le choix de la statistique locale des tableaux croisés
- Le choix (réduit) de la statistique globale : Khi2 (par défaut), CnC
- La couleur des graphiques

Pas plus que pour les ordinales le nombre de lignes et de colonnes ne sont contrôlées (sauf à dépasser 100).

### *$# ANOVA*

L'ANOVA (Analysis Of Variance) correspond au cas de figure tout à fait particulier du croisement mixte, c'est-à-dire au croisement d'une variable catégorielle (la nominale, ordinale ou logique en premier lieu) et d'une variable numérique ou assimilée.

L'enjeu de la démarche est le suivant : considérant les différences mesurées entre les individus selon la seconde variable, peuvent-elles être attribuées plutôt au « bruit » des caractéristiques individuelles des sujets, ou à leur appartenance à l'une ou l'autre des catégories selon la première variable ?

La représentation tabulaire est singulière, puisqu'on n'y trouve plus le recensement des combinaisons possibles des deux variables, mais plutôt, pour chaque catégorie de la première variable et pour l'ensemble, l'effectif, la moyenne et l'écart-type.

Analyse de la variance de E NOTE selon les positions de OPTION

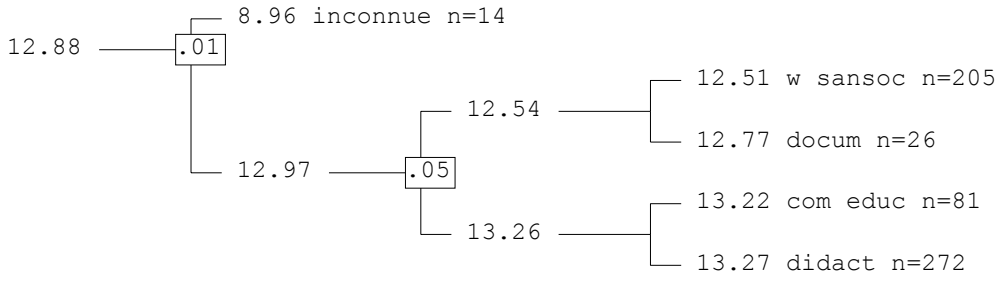
| Classe | Effectif | Moyenne | Ecart-type |
|---|---|---|---|
| com educ | 81 | 13.22 | 2.87 |
| didact | 272 | 13.27 | 3.50 |
| docum | 26 | 12.77 | 3.03 |
| inconnue | 14 | 8.96 | 5.77 |
| w sansoc | 205 | 12.51 | 3.31 |
| ENSEMBLE | 598 | 12.88 | 3.48 |

La statistique globale correspondante, unique, est le F de Snedecor-Fisher :

```
Variation totale : 7246.84 pour 597 degrés de liberté
  dont variation inter-classes : 293.38 pour 4 d.d.l., moyenne 73.35
    et variation intra-classes : 6953.46 pour 593 d.d.l., moyenne 11.73
F(4,593) = 6.26, s. à .0001
```

NB : à partir d'Octobre 2011, la présentation du détail du calcul de F est légèrement différente, et surtout ajoute une statistique %exp, pourcentage de variance expliquée, qui est le quotient de la variation inter-classe sur la variation totale. Cette statistique, parfois notée $\eta^2$, s'interprète de la même façon que le carré d'une corrélation.

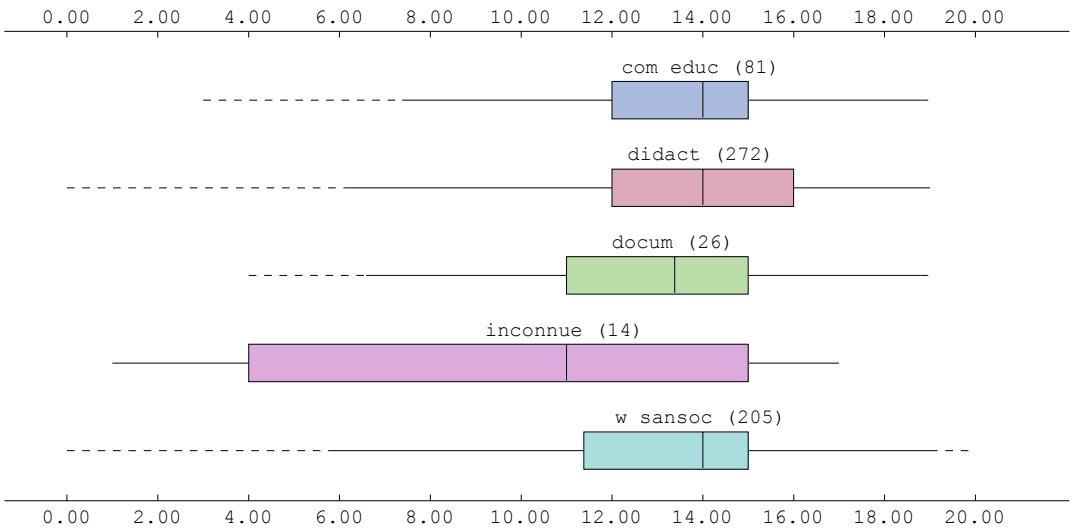
Exceptionnellement, la statistique locale est ici un graphique placé hors de la représentation tabulaire, qui expose une arborescence binaire de segmentation des catégories : celle-ci étant d'abord classées par moyennes croissantes, le logiciel cherche où placer une coupure qui opposera deux sous-ensemble les plus contrastés possible. Le critère est celui du |t| de Student pour des échantillons indépendants, analogue dans son usage au F de Snedecor-Fisher, mais pour seulement deux catégories à la fois.




La quantité figurant dans chaque petit cadre est le seuil de signification du |t| correspondant à la bifurcation. Ensuite chaque partie est à nouveau subdivisée selon le même principe, jusqu'à arriver aux catégories de la première variable. Les bifurcations qui ne portent pas ce petit cadre ont un |t| non significatif : la variable dont on étudie la variance ne permet pas de distinguer ces groupes. Le schéma ci-dessus pourrait s'interpréter ainsi : la catégorie {inconnue}, moyenne 8.96, contraste au seuil de .01 avec l'ensemble des autres catégories, moyenne 12.97, ensemble qui se subdivise au seuil de .05 en deux groupes : {w sanssoc} et {docum}, moyenne 12.54 , {com educ} et {didact}, moyenne 13.26. Le |t| ne permet pas d'aller plus loin dans les subdivisions.

L'arborescence des contrastes constitue une exception à la règle « si la statistique globale n'est pas significative, passe ton chemin », parce qu'il s'agit en fait non pas d'une statistique locale, mais d'une statistique globale sur d'autres variables virtuelles, celles qu'on obtiendrait en regroupant les valeurs de la variable catégorielle pour optimiser le contraste. *De ce fait, l'arborescence des contrastes peut être interprétée, même si le F n'est pas significatif.*

La représentation graphique correspondante est le schéma en « boîtes et moustaches » :




Pour chaque catégorie définie par la première variable, le schéma fait figurer un rectangle cadré par le premier et le troisième repère inter-quartile de la seconde variable et recoupé par la médiane. De part et d'autres du rectangle, les droites (moustaches) couvrent l'étendue de la distribution. En d'autres termes, entre le début de chaque ligne et le bord gauche de sa boîte, il y a environ 25% des sujets relevant de la catégorie ; entre ce bord gauche et le trait médian, encore 25% ; 25% aussi entre le trait médian et le bord droit de la boîte ; 25% enfin entre le bord droit de la boîte et la fin de la ligne. L'éventuelle partie pointillée des lignes signale la présence de sujets au-delà de deux écarts-types de part et d'autre de la moyenne, et donc une suspicion de valeurs rares ou aberrantes, qu'on peut négliger dans une description synthétique.

La représentation graphique de l'ANOVA est affectée par les options de choix des couleurs, ainsi que par une case à cocher [boîtes à moustaches polychromes] : si celle-ci n'est pas cochée, les boîtes ont toutes la même couleur.

## Tris de plus de deux variables

Hector permet d'évoquer jusqu'à quatre variables dans un même tri. Au-delà de deux, l'aspect des résultats dépend du type des deux dernières variables. S'il y a trois variables, la première ne peut être que d'un type générant des catégories (nominal, ordinal, logique) ; c'est le cas des deux premières s'il y a quatre variables.

Le principe général est que le croisement des deux dernières variables est réalisé pour chaque catégorie de la première (ou combinaison des deux premières). On obtient en quelque sorte un tableau de tableaux. Tous les éléments tabulaires, statistiques et graphiques sont réitérés pour chaque catégorie de la variable de contrôle.

Cette règle, qui est vraie quand les deux dernières variables sont du même type ou assimilé (croisement homogène), est un peu modifiée quand les deux dernières variables sont différentes (croisement mixte) : avec 3 variables, c'est l'ANOVA à deux facteurs ; avec quatre variables on revient à la règle commune.

En résumé, et en appelant x et y la première et l'éventuelle seconde variable (quand il y en a quatre), on peut rencontrer les cas suivants, étant entendu que x et y ne peuvent être que des variables catégorielles ($§£) :

- x## : croisement de numériques selon chaque catégorie de la variable x
- x££ : croisement de logiques selon chaque catégorie de la variable x

- x§§ : croisement d'ordinales selon chaque catégorie de la variable x
- x$$ : croisement de nominales selon chaque catégorie de la variable x
- x$# : ANOVA à deux facteurs
- xy## : croisement de numériques selon chaque combinaison de valeurs de x et y
- xy££ : croisement de logiques selon chaque combinaison de valeurs de x et y
- xy§§ : croisement d'ordinales selon chaque combinaison de valeurs de x et y
- xy$$ : croisement de nominales selon chaque combinaison de valeurs de x et y
- xy$# : ANOVA à deux facteurs selon chaque catégorie de la variable x

Il eût été théoriquement possible dans ce dernier cas de proposer une ANOVA à trois facteurs, mais les difficultés d'interprétation sont telles que cela n'en vaut sans doute pas la peine. De manière plus générale, ces tableaux de tableaux sont à employer dans des circonstances bien particulières, et avec parcimonie, parce qu'ils génèrent facilement des documents encombrants et finalement peu utiles. Une exception peut être faite pour l'ANOVA à deux facteurs, qui offre un graphique original.

*Anova à deux facteurs*

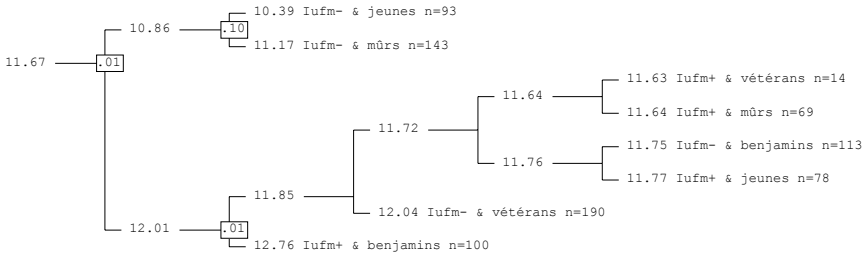
Analyse de la variance de moyenne générale selon IUFM et classe d'âge

| Classe 1 | Classe 2 | Effectif | Moyenne | Ecart-type |
|---|---|---|---|---|
| Iufm- | vétérans | 190 | 12.04 | 2.51 |
| | mûrs | 143 | 11.17 | 2.82 |
| | jeunes | 93 | 10.39 | 3.11 |
| | benjamins | 113 | 11.75 | 2.62 |
| Iufm+ | vétérans | 14 | 11.63 | 3.70 |
| | mûrs | 69 | 11.64 | 2.73 |
| | jeunes | 78 | 11.77 | 2.14 |
| | benjamins | 100 | 12.76 | 2.11 |
| TOUS | ENSEMBLE | 800 | 11.67 | 2.71 |

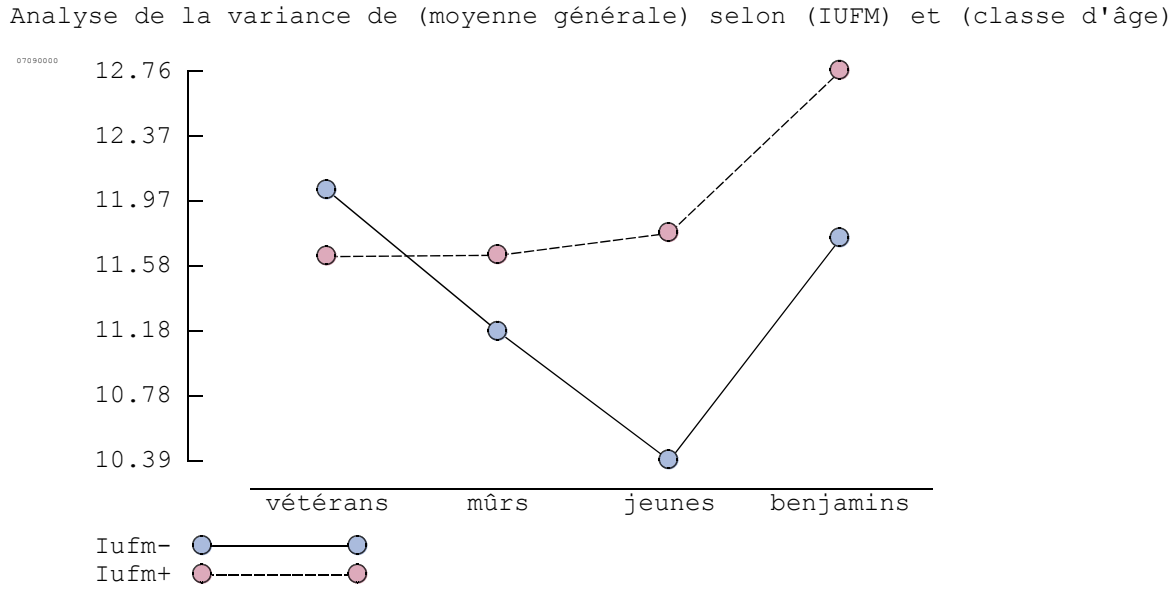
```
Variation totale : 5857.78 pour 799 degrés de liberté
  dont variation liée à v. n°1 : 74.22 pour 1 d.d.l., moyenne 74.22
  dont variation liée à v. n°2 : 186.59 pour 3 d.d.l., moyenne 62.20
  dont variation d'interaction : 74.93 pour 3 d.d.l., moyenne 24.98
    et variation intra-classes : 5522.04 pour 792 d.d.l., moyenne 6.97
Variable n°1 : F(1,792) = 10.64, s. à .01
Variable n°2 : F(3,792) = 8.92, s. à .01
 Interaction : F(3,792) = 3.58, s. à .05
```

La nouvelle présentation ajoute également la statistique %exp pour chaque facteur.

Le tableau ressemble à celui de l'Anova, mais deux variables catégorielles (les facteurs) sont mobilisées pour expliquer les variations de la troisième. On analyse donc l'influence de chacune des variables catégorielles prises séparément, et l'éventuel supplément d'influence liée à leur interaction, c'est-à-dire au fait qu'elles agissent ensemble. Le schéma correspondant à la statistique locale s'emploie, comme dans l'Anova à un facteur, à segmenter les cas de figure en recherchant les contrastes, mais les cas de figure élémentaires sont maintenant le résultat de combinaisons de valeurs de la première variable et de la seconde. Ainsi, les mieux classés du schéma suivant sont les 100 sujets qui sont à la fois des {Iufm+} et des {benjamins}. En queue de peloton, faiblement différenciés .10), les {Iufm-} d'âge moyen (ni benjamins, ni vétérans).



L'Anova à deux facteurs possède aussi un graphique spécifique, destiné à illustrer les interactions :




La première variable catégorielle fournit les deux « courbes » (en fait, pas vraiment des courbes, mais un simple chaînage qui réunit les points relevant de la même valeur de la première variable), la seconde les positions horizontales : si le résultat ne convient pas, on peut les permuter. C'est encore plus intéressant quand la seconde catégorielle est ordinale ou suspecte de l'être. La troisième variable, numérique, fournit l'échelle verticale, et les points sont placés à la hauteur correspondant à la moyenne du groupe considéré : le premier point bleu en haut et à gauche repère les 12.04 de moyenne des {Iufm-}{vétérans}.

L'interprétation du graphique d'interaction repose sur quelques principes simples : si les deux courbes sont à peu près parallèles, il n'y a sans doute pas d'interaction. Si l'une est plate et l'autre ascendante, ou descendante, il peut y avoir interaction. Si les courbes se croisent, il peut y avoir interaction contradictoire : c'est le cas ici, avec les {vétérans} qui réussissent mieux quand ils sont {Iufm-}, alors que c'est l'inverse dans les autres cas. Dans toutes ces situations, il faut avant tout se fier au seuil de signification fourni par les tests F de Snedecor-Fisher : nous avons ici une interaction contradictoire (ou non-ordinale) assez significative (.05).

## *Les filtres*

Un filtre est une restriction temporaire du corpus, du moins en ce qui concerne les sujets. Par exemple, on dispose d'un ensemble de données sur des personnes de différentes nationalités, et on souhaite exécuter une série de travaux statistiques, mais sur le sous-ensemble des étrangers seulement. On a donc besoin d'un procédé permettant, quelque tableau ou graphe qu'on demande, de ne travailler que sur ce sous-ensemble de la population.

*Définition*

Le procédé proposé par Hector est celui du filtre, qui est en l'occurrence une variable de type logique, à laquelle on fait jouer ce rôle de sélecteur. La variable logique doit exister avant d'être posée en filtre, mais elle n'a pas besoin d'être une variable d'origine : elle peut être une variable formulée, fabriquée pour l'occasion et détruite ensuite. L'accès aux outils de formulation est si aisé qu'il serait dommage de s'en priver. Ainsi, si on a une variable {nationalité}, avec une liste de nationalités dont {français}, on écrira une formule du genre de :

```
£ étranger
: nationalité <> français ;
```

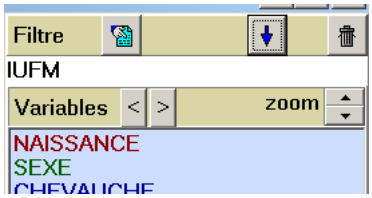
Pour la syntaxe détaillée du langage de dérivation, voir le chapitre qui lui est consacré. La formule ci-dessus parle d'elle même.

*Mise en oeuvre*

Dans le corpus qu'on a utilisé jusqu'à présent pour les exemples, on pourrait souhaiter ne travailler momentanément que sur les sujets dont la discipline d'origine relève des Lettres ou des Sciences Humaines. On prépare une variable logique *ad hoc* :

```
£ Lettres_&_Sciences_Humaines
: DISCIPLINE   in { let.lang sc.hum } ;
```

Au dessus de la liste générale des variables, un petit panneau sert à gérer les filtres :

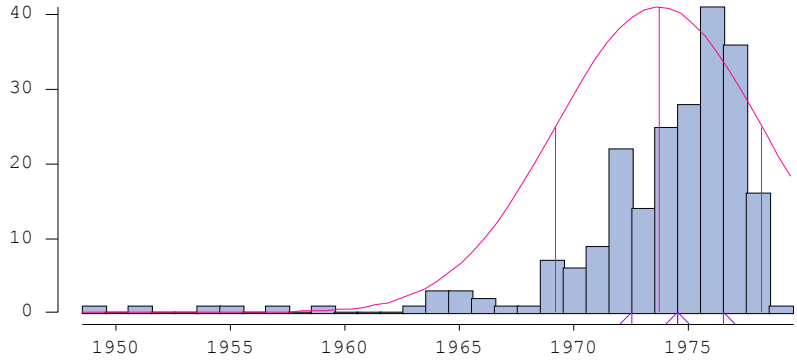



Ici, la variable {IUFM} est installée en position de filtre. Pour ce faire, on l'a sélectionnée dans la liste des variables, puis on a cliqué la flèche bleue, qui a envoyé son nom dans le cartouche placé en dessous.

Cette manœuvre ne semble pas avoir d'effet immédiat.

Cependant, si on demande un traitement quelconque, il s'effectue sous le contrôle du filtre :

```
NAISSANCE
sous le filtre Lettres & Sciences Humaines
Valeur modale : 1976 (n=41)
Médiane entre 1974 & 1975
Moyenne 1973.67, écart-type 4.52
```



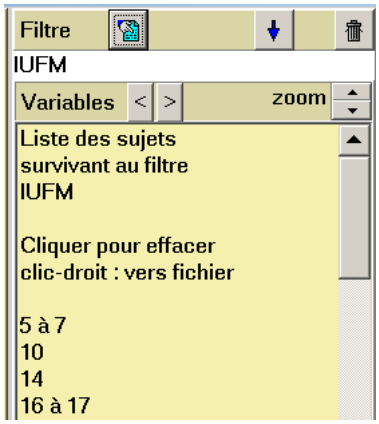

Ce contrôle est d'ailleurs rappelé sous l'intitulé du tri, pour que l'usager n'oublie pas qu'il a posé un filtre. Le tri demandé ici est très simple, mais la technique du filtre fonctionne aussi avec les tris les plus complexes : à la limite, on peut l'utiliser comme une manière d'augmenter le nombre de variables triées simultanément.

Le filtre peut être aussi complexe que l'on voudra : il suffit de fabriquer plusieurs filtres élémentaires, et de les assembler dans une autre formule de dérivation. Si l'on voulait par exemple se concentrer uniquement sur les {Lettres & Sciences Humaines} qui se destinent à l'Iufm, on peut construire ceci :

```
£ Lettres-Sces_Hum._et_IUFM
: Lettres_&_Sciences_Humaines et Iufm ;
```

*Survivants*

Le bouton orné d'une main qui pointe dans une liste sert à afficher les numéros d'ordre des survivants, c'est-à-dire des sujets qui satisfont aux conditions du filtre.




Ceci est particulièrement utile dans la phase de nettoyage d'un corpus après saisie et de recherche d'erreurs. Supposons qu'on ait enquêté auprès d'une population d'adolescents, et donc des 13-17 ans (ou toute autre définition qu'on se donnera de cette tranche d'âge). Or, au tri sur les âges, on trouve un sujet qui affiche 46 ans. On subodore qu'il s'agit d'une erreur, et on souhaite la vérifier et éventuellement la corriger. Dans tous les cas, on a besoin d'identifier le sujet en cause. Une solution coûteuse consiste à réviser visuellement les données ; l'autre consiste à poser un filtre temporaire qu'on rédigera ainsi :

```
£ erreur : Age = 46 ;
```

L'appel aux survivants fournira le numéro du coupable, et lui seul. Il sera alors aisé de vérifier dans les documents-source, et éventuellement de corriger avec l'éditeur de données.

Si l'on sélectionne une nouvelle variable logique comme filtre, elle se substitue à l'ancienne, puisqu'il ne peut y avoir qu'un filtre à la fois.

Le bouton orné d'une poubelle sert à vider la boîte à filtre, après quoi les sujets du corpus sont restaurés dans leur intégralité.

## *Exporter les résultats*

Les tableaux, graphiques et statistiques qui sont affichés dans la page TRAITEMENT peuvent être envoyés vers l'extérieur, et typiquement vers un traitement de texte, au moyen du bouton initialement porteur d'une flèche et de la mention « document à définir », mention remplacée après usage par le nom du fichier qui a été désigné pour servir de document, c'est-à-dire d'intermédiaire, de moyen de transport entre Hector et un traitement de texte.

C'est par ce type de procédé que les résultats ont été importés dans le présent document. Le format du fichier dépend du système d'exploitation sous lequel tourne Hector.

L'usage de ce procédé et les moyens de le paramétrer judicieusement sont décrits dans le Chapitre « Exploitation des résultats ». On peut quand même noter qu'il est possible de travailler dans Hector avec un traitement de texte ouvert pour récupérer immédiatement les résultats, mais qu'il est nécessaire pour ce faire de fermer le document, en droite-cliquant le bouton où est affiché son nom ; en effet, le document est un fichier qui ne saurait être utilisé par deux applications à la fois, Hector qui l'écrit et le traitement de texte qui veut le lire.

# TRAITEMENTS COLLECTIFS

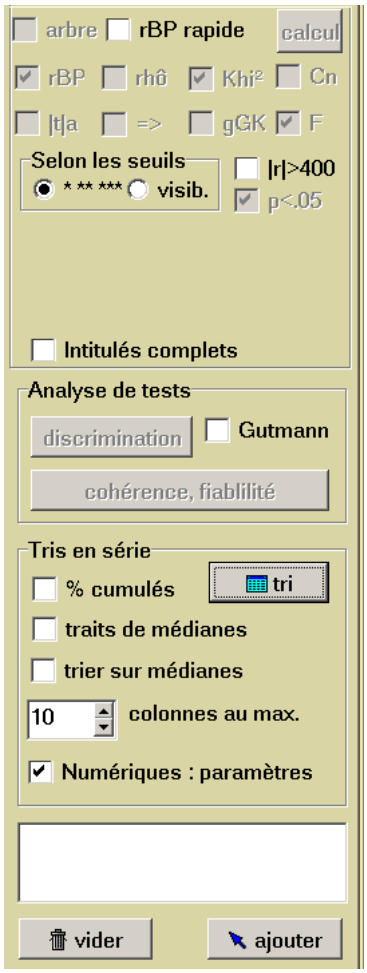
Les traitements collectifs sont dits tels parce qu'ils agissent sur des collections dans leur ensemble plutôt que sur des variables isolées.

---

## *La page TRAITEMENTS, volet [collectifs]*

En bas du panneau de commande de la page TRAITEMENTS, on trouve un système d'onglets :

Si on sélectionne l'onglet du volet [collectifs], le panneau prend l'aspect suivant :




Mis à part les organes connus, on repère, de haut en bas :

- un cadre [Matrice de statistiques], contenant :
  - une case à cocher [arbre]
  - une case à cocher [rBP rapide]
  - un bouton [calcul]
  - des cases à cocher évoquant des tests statistiques

- un sélecteur [selon les seuils]
- une case à cocher [|r|>400]
- une case à cocher [p<.05]
- une case à cocher [intitulés complets]

- un cadre [analyse de tests], contenant
  - un bouton [discrimination]
  - une case à cocher [Gutmann]
  - un bouton [cohérence, fiabilité]
- un cadre [Tris en série], contenant :
  - une case à cocher [% cumulés]
  - un bouton [tri]
  - une case à cocher [traits de médiane]
  - une case à cocher [trier sur médianes]
  - un compteur numérique [nombre max colonnes]
  - une case à cocher [Numériques : paramètres]
- Une zone vague comportant
  - une boîte à collections sélectionnées, avec deux places
  - des boutons [ajouter], [vider]

Trois sortes différentes de traitements peuvent être effectués dans ce contexte : des matrices de statistiques, des analyses de tests et des tris en série. Commençons par le dernier cas, qui est aussi le plus simple.

---

## *Tris en série*

L'idée de base des tris en série est que des variables qui se ressemblent, qui possèdent soit les mêmes valeurs de texte (nominales, ordinales, et implicitement logiques), soit des valeurs de nombres compatibles (numériques, calendaires) peuvent, plutôt qu'être triées une par une dans le volet [tris], avec une succession fastidieuse de clics de souris, faire l'objet d'un tri global, d'un seul coup, et de préférence dans un tableau commun économisant les en-têtes redondantes.

Ce souhait peut notamment intervenir dans les phases de vérification du corpus, où l'on doit procéder à des tris de vraisemblance qui imposent de ne laisser aucune variable dans l'ombre.

On commence par installer un nom de collection dans la boîte des collections sélectionnées, soit en la double-cliquant dans sa liste générale des collections, soit en la sélectionnant et en cliquant le bouton [ajouter].

Si la boîte comporte déjà une ou des collections, le bouton [vider], ou un droite-clic dans la liste générale, videra la boîte.

Peu importe qu'il y ait une ou deux collections dans la boîte de sélection : seule la première sera prise en compte.

Cliquer le bouton [trier] peut produire ce genre de résultat :

Tri en série de la collection {z_notes}

| [classes[ | [-5.0 | | [-4.0 | | [-3.0 | |
|---|---|---|---|---|---|---|
| z_A_NOTE | 0 | 0% | 5 | 1% | 20 | 3% |
| z_B_NOTE | 0 | 0% | 0 | 0% | 23 | 4% |
| z_C_NOTE | 0 | 0% | 0 | 0% | 9 | 2% |
| z_D_NOTE | 0 | 0% | 5 | 1% | 23 | 4% |
| z_E_NOTE | 0 | 0% | 13 | 2% | 20 | 3% |
| z_O1_NOTE | 0 | 0% | 14 | 2% | 9 | 1% |

On n'a fait figurer ici qu'une partie de l'image.

Il s'agit d'une collection de variables numériques, dont les tris sont fournis par classes, de manière à ce que le système de classes soit à la fois valable pour toutes les variables de la collection, et compatible avec la valeur actuelle de [nombre max de colonnes].

Les pourcentages sont cumulés de gauche à droite parce que l'option [Pourcentages] est à [cumulés], et non à [simples]. L'option [cumulés] n'a d'effet que dans le cas d'une variable numérique ou calendaire.

On notera que les noms de variables, pour pouvoir rentrer dans un cadre habituellement réservé aux étiquettes (10 caractères), ont subi l'algorithme d'abréviation.

> Pour les variables numériques seulement, et à condition que les valeurs ne soient pas regroupées, la case à cocher [traits de médiane] a pour effet qu'un trait vertical est placé dans chaque ligne de manière à figurer le coupure médiane, telle que les effectifs soient approximativement partagés par moitiés. De plus, si la case [trier sur médianes] est cochée, l'ordre des variables suit l'ordre croissant des médianes. Ce dispositif est spécialement utile dans le cas de valuateurs (variables numériques de 1 à 8 reflétant des opinions ou représentations), car plus la médiane est à gauche et plus l'opinion correspondante est rejetée, et plus elle est à droite plus elle est acceptée : on obtient donc directement un classement des valuateurs par adhésion croissante.

Si la case [Numériques : paramètres] est cochée, le tri en série de variables numériques affichera pour chaque variable la moyenne, l'écart-type et la médiane.

Dans le cas de variables de type texte (nominales ou ordinales), l'unicité du tableau n'est pas garantie, parce que les étiquettes sont différentes :

Tri en série de la collection ords

| | insuff | | médiocre | | satisf | | supérieur | | Total |
|---|---|---|---|---|---|---|---|---|---|
| classe de résultats | 201 | 25% | 191 | 24% | 203 | 25% | 205 | 26% | 800 |

| | vétérans | | mûrs | | jeunes | | benjamins | | Total |
|---|---|---|---|---|---|---|---|---|---|
| classe d'âge | 246 | 25% | 272 | 28% | 216 | 22% | 233 | 24% | 967 |

Dans ce cas, seules des variables possédant les mêmes jeux d'étiquettes pourraient faire tableau commun. En revanche, des variables logiques ont implicitement les mêmes étiquettes :

Tri en série de la collection logs

| | Faux | | Vrai | | Total |
|---|---|---|---|---|---|
| CHEVAUCHE | 937 | 97% | 30 | 3% | 967 |
| IUFM | 685 | 71% | 282 | 29% | 967 |
| FINIR | 316 | 33% | 651 | 67% | 967 |

## *Matrices de statistiques*

Les matrices de statistiques proviennent du constat suivant : à certaines étapes du travail d'analyse d'un corpus, ce dont on a besoin n'est pas tant la série détaillée de croisements de variables prises

deux à deux qu'une vision synthétique des relations que des ensembles complets de variables entretiennent entre eux. Par exemple, on dispose d'une collection de n variables numériques représentant des mesures, et on souhaite connaître la corrélation de chacune de ces variables avec chacune des autres. Pour obtenir ce résultat dans le volet [tris] de la page TRAITEMENTS, il faudrait opérer n*(n-1)/2 croisements et relever à la main le résultat du test r de Bravais-Pearson pour garnir un tableau.

Dans le volet [collectifs], ce service est accessible directement :

- on installe deux collections dans la boîte de sélection
- on sélectionne le ou les tests souhaités parmi ceux qui sont disponibles
- on clique le bouton [calcul]

Le résultat s'affiche ainsi :

Collection ords × collection notes

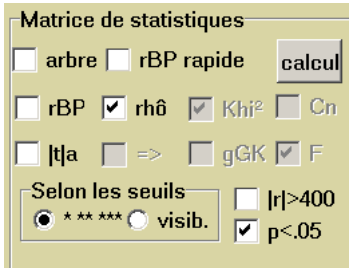
Matrice des coefficients de corrélation par rangs rhô de Spearman

| | ANote | BNote | CNote | DNote | ENote | O1Note | O2Note | O3Note |
|---|---|---|---|---|---|---|---|---|
| ClasDeRésu | 0.609 *** | 0.583 *** | 0.607 *** | 0.552 *** | 0.477 *** | 0.589 *** | 0.580 *** | 0.484 *** |
| ClasseDÂge | -0.067 | 0.007 | -0.038 | 0.090 ** | 0.069 | -0.003 | 0.153 *** | 0.054 |

Matrice des F de Snedecor-Fisher

| | ANote | BNote | CNote | DNote | ENote | O1Note | O2Note | O3Note |
|---|---|---|---|---|---|---|---|---|
| ClasDeRésu | 120.363 *** | 108.075 *** | 114.094 *** | 101.149 *** | 79.761 *** | 89.079 *** | 114.088 *** | 66.383 *** |
| ClasseDÂge | 10.687 *** | 4.758 *** | 8.497 *** | 2.216 * | 3.401 ** | 1.441 | 7.882 *** | 1.593 |

Pourquoi y-a-t-il deux tableaux ? Parce que deux types de tests étaient disponibles et ont été utilisés :



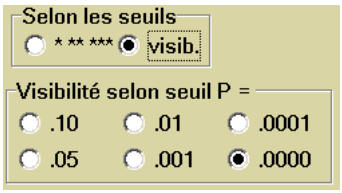

On avait en effet installé dans la boîte à sélection deux collections, {ords} qui est une collection de variables ordinales, et {notes} qui est une collection de variables numériques. A noter que la

collection numérique se met forcément en seconde place, comme dans la page des tris.

Les règles qui gouvernent la disponibilité des tests lors des croisements s'appliquent aussi ici : le croisement ordinale × numérique autorise le rhô de Spearman et le F de Snedecor-Fisher. Les tests non autorisés sont restés en grisé, cochés ou non.

Si l'on retourne examiner les tableaux, on constate que certaines valeurs des tests sont accompagnées d'étoiles : celles-ci codent le seuil de significativité des tests (la probabilité de l'erreur), à raison de trois étoiles pour un seuil P=.01 ou moins, deux étoiles pour un seuil P=.05, une étoile pour P=.10, et aucune étoile pour des relations non significatives.

Dans le cadre [Matrice de statistiques] existe une case à cocher [p<.05]. Si elle est cochée, les tests non significatifs ou seulement à .10 (faiblement significatifs) ne sont pas affichés, ce qui permet de repérer plus rapidement les relations intéressantes.




Si le sélecteur [Selon les seuils] est à « visib. », pour visibilité, plus d'étoile, mais l'affichage sélectionne les valeurs significatives au seuil qui a été choisi.

Collection ords × collection notes
Matrice des coefficients de corrélation par rangs rhô de Spearman
au seuil P = .001

| | ANote | BNote | CNote | DNote | ENote | O1Note | O2Note | O3Note |
|---|---|---|---|---|---|---|---|---|
| ClasDeRésu | 0.609 | 0.583 | 0.607 | 0.552 | 0.477 | 0.589 | 0.580 | 0.484 |
| ClasseDÂge | | | | | | | 0.153 | |

Matrice des F de Snedecor-Fisher
au seuil P = .001

| | ANote | BNote | CNote | DNote | ENote | O1Note | O2Note | O3Note |
|---|---|---|---|---|---|---|---|---|
| ClasDeRésu | 120.363 | 108.075 | 114.094 | 101.149 | 79.761 | 89.079 | 114.088 | 66.383 |
| ClasseDÂge | 10.687 | | 8.497 | | | | 7.882 | |

Dans l'exemple ci-dessus, on a « croisé » deux collections différentes, mais on peut aussi croiser une collection avec elle-même (ici la collection {notes} ) :

Collection notes × elle-même
Matrice des coefficients de corrélation r (Bravais-Pearson)
au seuil P = .0000

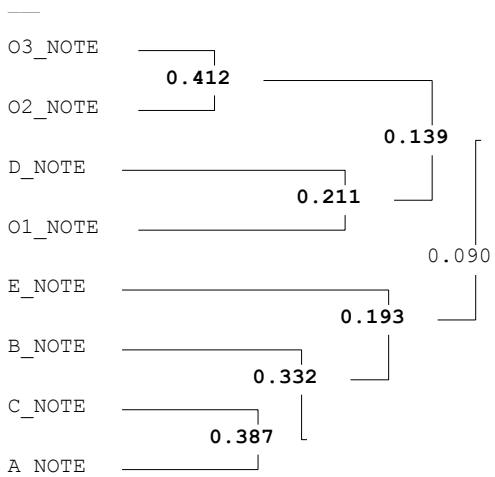
| | ANote | BNote | CNote | DNote | ENote | O1Note | O2Note | O3Note |
|---|---|---|---|---|---|---|---|---|
| ANote | | 0.353 | 0.387 | 0.318 | | 0.278 | 0.307 | |
| BNote | 0.353 | | 0.332 | 0.242 | 0.257 | 0.226 | 0.222 | |
| CNote | 0.387 | 0.332 | | | 0.214 | | 0.317 | 0.211 |
| DNote | 0.318 | 0.242 | | | | 0.211 | 0.207 | |
| ENote | | 0.257 | 0.214 | | | | 0.196 | |
| O1Note | 0.278 | 0.226 | | 0.211 | | | 0.274 | |
| O2Note | 0.307 | 0.222 | 0.317 | 0.207 | 0.196 | 0.274 | | 0.412 |
| O3Note | | | 0.211 | | | | 0.412 | |

On a utilisé l'option de visibilité au seuil le plus exigeant : avec les astérisques, on aurait eu des *** partout. On constate aussi une série de cases vides sur la diagonale : la corrélation d'une variable avec elle-même est 1.000 par définition, ce n'est donc pas un résultat intéressant. Ici la matrice est symétrique autour de la diagonale, mais c'est une propriété du test utilisé (r BP). D'autres tests, comme F et l'implication =>, ne sont pas symétriques, c'est-à-dire que le résultat du test appliqué à la paire de variables X et Y n'est pas nécessairement le même qu'appliqué à la paire Y et X. Dans le cas de F, il n'est même pas applicable puisqu'il s'agit d'un test entre variables d'un type différent.

L'arborescence des parentés

Près du bouton [calcul] des matrices de statistiques, dans l'onglet [collectifs], on trouve une case à cocher [arbre]:

Si on demande le calcul d'une matrice de corrélation (r de Bravais –Pearson ou rhô de Spearman), on obtient en plus un schéma de ce genre :




Cela ressemble vaguement à un arbre couché, dont le tronc serait à droite, et dont les feuilles, à gauche, sont les items.

La démarche de construction de cette arborescence consiste à rechercher les deux items les mieux corrélés, et à les assembler : ici les items O3NOTE et O2NOTE. Assemblés, il constituent une entité dont on détermine la corrélation avec les autres entités, dont les items encore isolés. Cette corrélation avec la nouvelle entité est, par construction, la plus petite corrélation constatée avec un des éléments de cette entité.

Cela a pour conséquence que la corrélation portée au sommet d'un groupe d'items est en quelque sorte la corrélation minimale garantie entre deux quelconques des membres du groupe.

Les corrélations écrites en **gras** sont significatives à .01, celles écrites en caractères ordinaires sont significatives à .05, celles écrites en *italique* sont significatives à .10 ou pas du tout.

Il s'agit donc de mettre en évidence la structure sous-jacente des parentés entre variables.

Deux options concernent spécifiquement les matrices de corrélation : si la case [rBP rapide] est cochée, l'algorithme utilisé pour le calcul des corrélations et beaucoup plus rapide, mais il offre une particularité : il ne prend en compte que les sujets qui ont une valeur définie pour *toutes* les variables de la collection, alors que l'algorithme ordinaire traite les variables couple par couple, avec la conséquence que, s'il y a des valeurs non définies, les corrélations ne sont pas toutes calculées exactement sur les mêmes effectifs. Par ailleurs, si la case [|r|>400] est cochée, seules les corrélations égales ou supérieures à .400 sont affichées, ceci pour faciliter la lecture de grandes matrices de variables bien corrélées entre elles.

# *Analyse de tests*

Cet ensemble de fonctions est destiné à la vérification des qualités métriques d'épreuves ou de tests. Il se révèle particulièrement utile en psychométrie et en docimologie, comme pour toutes les professions qui, comme les Orthophonistes, font un grand usage des tests.

## Discrimination et difficulté

Une collection logique ou numérique avec uniquement des valeurs 0/1 étant sélectionnée, l'appui sur le bouton [discrimination] permet d'obtenir ce genre d'affichage :

```
Analyse de la collection logique ou binaire {tous les items}
taux de réussite, difficulté
indice de discrimination et qualité

_____________________________________________________________
REUSSITE   faibles  médians   forts   DISCRIM  qualité   item
```

```
0,26  -      0,11     0,25     0,51     0,40     Ok       a
0,82  +      0,63     0,89     0,93     0,30      -       b
0,62  =      0,24     0,71     0,95     0,70     Ok**     c
0,52  =      0,23     0,53     0,91     0,68     Ok**     d
0,29  -      0,10     0,22     0,73     0,63     Ok**     e
0,18  -      0,09     0,14     0,40     0,32     Ok       f
0,22  -      0,07     0,16     0,56     0,48     Ok*      g
0,63  =      0,25     0,72     0,96     0,71     Ok**     h
0,52  =      0,15     0,59     0,86     0,71     Ok**     i
0,70  +      0,45     0,75     0,92     0,47     Ok*      j
```

```
La première colonne est celle du taux de réussite. Il est exprimé ici comme une
fréquence entre 0 et 1, qui équivaut à l'intervalle 0% à 100%.
```

Le système de signes qui commente le taux de réussite est arbitrairement inspiré de l'expérience :

- -- de 0 à 15%
- - de 16 à 35%
- = de 36 à 64%
- + de 65 à 84%
- ++ de 85 à 100%

Quelle est l'importance des taux de réussite ? Eh bien un item très difficile ou très facile n'est pas très intéressant du point de vue de l'information qu'il peut apporter, car l'analyse statistique vise à faire apparaître des différences entre les cas et les situations ; de ce point de vue, les taux de réussite les plus intéressants en termes de théorie de l'information sont autour de 50% : ce sont ceux qui font faciliter les tâches de comparaison et de classement.

Un test dont tous les items seraient trop difficiles ou trop faciles n'est tout simplement pas adapté à la population étudiée. On notera au passage (on y reviendra) que la difficulté d'un item n'a de sens que vis-à-vis d'une population donnée : ce qui est difficile pour vous ne l'est pas nécessairement pour moi, et *vice versa*.

Les trois colonnes suivantes sont encore des taux de réussite, mais calculés pour des sous-groupes distincts d'individus : les *faibles*, les *médians* et les *forts*.

Comment sont constitués ces groupes ? Sur la base du score obtenu en totalisant les scores obtenus aux items. Les faibles ont les 27% de plus faibles au regard de ce score total, les forts sont les 27% de plus forts, et les médians les 46% qui restent au milieu.

Ainsi, pour l'item *a*, le taux de réussite des faibles est de 11%, celui des médians de 25%, celui des forts de 51%.

L'intérêt principal de la manœuvre réside dans la cinquième colonne, qui contient l'indice de discrimination de chaque item : c'est la différence entre le taux de réussite des forts et celui des faibles.

Un indice de discrimination élevé dénote un item utile dans la mesure où l'on souhaite construire un test qui sépare clairement les individus selon des niveaux de performance contrastés. A l'inverse, un item qui est plus ou moins réussi, mais à peu près autant par les faibles et les forts, ne traite probablement pas des mêmes compétences que les autres items de l'épreuve.

La colonne qualité du tableau expose des appréciations sur la qualité discriminante des items, de manière à permettre le repérage rapide des anomalies :

- Ok++ correspond à une discrimination d'au moins 0,50, qu'on pourrait qualifier d'excellente
- Ok+ va de .40 à .49, c'est très satisfaisant
- Ok va de .30 à .39, c'est correct

- - va de .20 à .29, c'est faible : l'item n'est à conserver que si on ne peut absolument pas s'en passer (par exemple parce qu'on s'est laissé enfermer dans une situation où on n'en n'a pas de rechange).
- -- va de .10 à .19 : il est préférable d'éviter d'utiliser un tel item
- ?? est pour les items dont l'indice de discrimination tombe en dessous de .10 : ceux-là n'apportent rien à l'épreuve
- !!! est pour les items aberrants, par exemple ceux que les faibles réussissent mieux que les forts.

Bien qu'il s'agisse d'un indice formel et insensible à la signification, il va de soi que des indices de mauvaises qualité remettent en cause la construction de l'épreuve et/ou son adaptation au public visé.

On prendra également garde au fait que les indices de discrimination sont plus faibles quand l'épreuve contient un nombre important d'items (20 ou 30), parce qu'avec un nombre plus petits d'items, la présence de l'item lui-même dans le total qui permet de définir les trois classes d'individus tend à biaiser les résultats en faveur de l'item ; dans le cas de petits nombres d'items, comme ici, il faut donc être plus exigeant.

*Modèle de Guttman*

Si la case mG est cochée, cela indique qu'on souhaite confronter la collection au modèle de Guttman.

Celui-ci s'applique normalement à une collection d'items binaires mesurant à des niveaux divers une même compétence, ou une même dimension de compétence. Le principe est que, si un sujet a réussi un item d'un certain niveau de difficulté, on s'attend à ce qu'il ait réussi aussi à tous les items de difficulté inférieure.

Ainsi, avec quatre items de difficulté croissante a, b, c, d, le profil de réussite 1 1 0 0 est conforme au modèle, puisqu'il est celui des individus qui ont réussi a et b, mais ni c ni d.

En revanche, un profil 1 0 1 0 n'est pas conforme, puisque les individus concernés n'ont pas réussi b, alors qu'ayant réussi c ils auraient « dû » réussir aussi b.

Les seuls profils acceptables sont les suivants :

0 0 0 0, 1 0 0 0, 1 1 0 0, 1 1 1 0 et 1 1 1 1

Ils ont en commun qu'aucun 1 ne doit apparaître à la *droite* d'un 0, et, réciproquement, aucun 0 à *la gauche* d'un 1.

Si l'on arrange les profils conformes au modèle dans un tableau approprié, on obtient le tableau suivant :

| a | b | c | d |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |

La forme en escalier justifie le nom de modèle pyramidal parfois donné à cette forme. Il a des chances de se produire dans un système ou chaque capacité plus rare englobe les précédentes.

Le coefficient de reproductibilité de Guttman est le quotient du nombre de cases convenables par le nombre total de cases (nombre de sujets x nombre de variables) dans le grand tableau à une ligne par sujet et une colonne par variable. Une case non convenable est une case qui contient un 0 (échec) à la gauche d'un 1 (réussite à un item réputé plus difficile). L'autre mesure est le pourcentage de sujets qui présentent des profils rigoureusement compatibles avec le modèle de

Guttman. La fréquence croissante des erreurs permet d'identifier dans quelle mesure chaque variable contribue au nombre total d'erreurs.

```
Analyse de la collection logique ou binaire pas de problèmes % phrases
taux de réussite, difficulté
indice de discrimination et qualité

Reproductibilité (modèle de Guttman) : 0,83
% de sujets rigoureusement conformes = 59,01 (théorique 31,25)
seuils de confiance à .01, .001, .0000 : 36,05 37,62 39,50
fréquence croissante des erreurs :
0,00 pas de pb % phrases
0,18 pas de pb % phrases exp
0,32 pas de pb % phrases m.f
0,51 pas de pb % phrases +2t
```

La reproductibilité de .83 est relativement importante, mais pas assez pour qu'on puisse se fier entièrement au modèle : on exige usuellement pour cela un coefficient de .90. La ventilation des erreurs par variable peut permettre à un stade de la mise au point du test, si le modèle pyramidal est souhaité, quel(s) item(s) il faudrait exclure pour améliorer le coefficient de Guttman.

Alors que le coefficient de Guttman fait plutôt porter la « responsabilité » des erreurs sur les items, le taux de sujets conformes mesure à quel point la distribution observée s'éloigne du taux théorique de sujets conformes sous l'hypothèse de réponses indépendantes. Ce n'est pas une mesure très exigeante en soi que l'éloignement du modèle aléatoire. Aussi utilise-t-on des seuils de décision très fins : .01, .001, .0000 (quasi certitude). La question à laquelle il est répondu n'est pas « Est-ce que cette collection présente une structure pyramidale ? », mais « A quel point le nombre de sujets conformes au modèle de la structure pyramidale s'écarte-t-il de ce que le hasard aurait pu provoquer ? »

Dans l'exemple, le taux de conformité supérieur à 59% permet d'écarter l'hypothèse nulle (aléatoire) : il y a bien une tendance à la structure pyramidale, mais elle n'est pas parfaite (coefficient de Guttman à .83). Un taux de conformité aussi significatif pourrait conduire à rechercher pour quelles parties de la population étudiée le modèle serait mieux satisfait… mais ceci nous éloigne du propos principal.

## Cohérence, fiabilité

La seconde vérification des qualités métriques des épreuves et tests concerne la cohérence d'une épreuve, ou sub-test. Le modèle sous-jacent est que l'épreuve est constituée d'items parallèles en contenu et en difficulté, qui constituent autant d'indicateurs imparfaits mais convergents d'une compétence sous-jacente, polluée par le « bruit » de la situation de test.

Cette notion de cohérence est extrêmement importante d'un point de vue pratique, car elle légitime le fait de procéder à des additions de scores d'items pour produire un score d'épreuve. Pour le dire plus familièrement, une cohérence élevée garantit qu'on additionne bien des poireaux avec des poireaux et non avec des carottes.

On y accède en cliquant le bouton [cohérence, fiabilité], la collection à étudier étant dans la boîte à étudier les collections.

```
Analyse de la collection numérique ou logique {tous les items}
```

| moyenne | écart-type | r(i,T-i) | item |
|---|---|---|---|
| 0,26 | 0,44 | 0,118 | a |
| 0,82 | 0,38 | 0,160 | b |
| 0,62 | 0,48 | 0,386*** | c |
| 0,52 | 0,50 | 0,312*** | d |
| 0,29 | 0,45 | 0,290*** | e |
| 0,18 | 0,38 | 0,127 | f |

```
0,22          0,41        0,234**       g
0,63          0,48        0,385***      h
0,52          0,50        0,301***      i
0,70          0,46        0,201*        j

Corrélations item/test (cohérence) : min, moy, max : 0,118, 0,251, 0,386
Alpha de Cronbach (fiabilité) = 0,572
```

La première colonne reprend le score moyen à l'item, qui équivaut au taux de réussite quand la variable est binaire (cette fonctionnalité est également accessible aux items non-binaires). La seconde colonne fournit l'écart-type de ce score (indice de dispersion).

La troisième colonne contient la véritable mesure de cohérence : la corrélation item-test. Elle est ainsi appelée dans le sens ou le test serait la seule superstructure à l'item, mais évidemment il s'agit d'une corrélation item-subtest ou item-épreuve.

Plus précisément, il s'agit, pour chaque item, de mesurer la corrélation entre l'item lui-même et la somme des items de l'épreuve, l'item lui-même exclu. C'est ainsi qu'il faut comprendre le titre un peu sybillin de la colonne : r(i,T-i) ; *r* est mis pour corrélation entre *i*, l'item et *T-i*, la somme des items, sans l'item considéré.

Quand cette corrélation est élevée, cela signifie que l'item est bien à sa place dans cette épreuve, qu'il contribue efficacement à constituer la mesure globale que sera la somme des scores aux items, autrement dit le score à l'épreuve : une forte cohérence légitime le fait même de calculer un tel score par addition de scores partiels.

Certaines corrélations portent une, deux ou trois astérisques *. Avec *, c'est une corrélation significative au seuil de .10, avec **, au seuil de .05, avec ***, au seuil de .01. Sans signe, c'est non significatif.

La dernière ligne présente la statistique alpha de Cronbach , qui est une mesure de fiabilité. Ce n'est pas une corrélation, mais une estimation de la probabilité que les items mesurent la même chose, que les erreurs se compensent pour que la somme délivre la mesure d'une valeur sous-jacente. Ici, des valeurs comme 0,572 ou 0,560 sont très insuffisantes : il n'y a pratiquement pas plus d'une chance sur deux qu'un tel modèle soit réaliste. Un alpha de 0,750 paraît un minimum, et 0,900 est très bon.

On peut améliorer la statistique de fiabilité alpha de Cronbach en augmentant le nombre d'items, à condition que ceux-ci soient au moins aussi cohérents que ceux qui existent déjà. La formule suivante :

$$k = [\ a_1\ (1 - a_0)\ ]\ /\ [a_0\ (1 - a_1)\ ]$$

où $a_0$ désigne l'alpha de Cronbach actuel, et $a_1$ l'alpha de Cronbach souhaité, fournit k, coefficient par lequel il faut multiplier le nombre actuels d'items pour espérer atteindre la fiabilité souhaitée (toujours sous la condition d'items cohérents).

En partant de l'épreuve {cdehi} et avec comme objectif une fiabilité à 0,750,

$k = [\ 0{,}75 \times (\ 1\text{- }0{,}56)\ ]\ /\ [\ 0{,}56 \times (\ 1 - 0{,}75)\ ]$, soit $k = 2{,}36$

Il faudrait donc passer à une douzaine d'items de même qualité pour atteindre une fiabilité minimale. Pour atteindre une fiabilité de 0,900 (90%), il faudrait multiplier le nombre d'items par 7 !

Une formule dérivée de celle-ci est utilisée pour calculer un Alpha comparable pour 10 items, c'est-à-dire l'alpha qu'on obtiendrait en ramenant le nombre d'items à 10, sous condition de conserver une même qualité de cohérence. Cette mesure, qui n'est en rien la vraie valeur de la cohérence, est présente uniquement pour permettre des comparaisons entre tests dotés d'un nombre différent d'items.

# LE PLAN DE PROJECTION

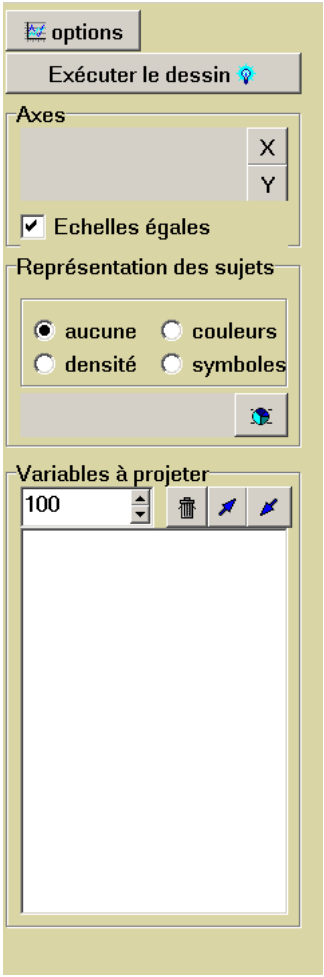
Le plan de projection ne constitue pas tant une méthode statistique qu'une technique de représentation de résultats. Si cette technique est régulièrement associée dans la littérature statistique classique aux analyses factorielles, Hector la généralise à un usage plus ouvert.

*Le plan de projections n'est accessible que dans la version professionnelle-recherche d'Hector. Sa preésence se manifeste par un troisième onglet en bas du panneau de commande des TRAITEMENTS.*

---

## *La page TRAITEMENTS, volet [projection]*

Si dans la page TRAITEMENTS on sélectionne l'onglet du volet [projection], ce dernier prend l'aspect suivant :




On distingue, de haut en bas :

- le bouton [options]
- le bouton [Exécuter le dessin]
- le cadre [Axes], avec :
  - la boîte de sélection de l'axe horizontal (X)

- la boîte de sélection de l'axe vertical (Y)
- la case à cocher [échelles égales]

- le cadre [Sujets], avec :
  - la boîte de choix de la représentation des sujets
  - la boîte de sélection de la boîte des « couleurs ou symboles selon ... »
- le cadre [Variables à projeter], avec :
  - un compteur pour le taux applicable aux pseudo-rayons des ellipses
  - les boutons pour vider, ajouter et enlever
  - la liste des variables à projeter

Le bouton [Exécuter le dessin] est celui qu'on actionne en dernier, quand tout est prêt. Le bouton [Options] donne accès au même panneau d'options que dans le volet [tris] ; il est utile ici pour les couleurs et pour les paramètres du nuage de densité, s'il est utilisé.

---

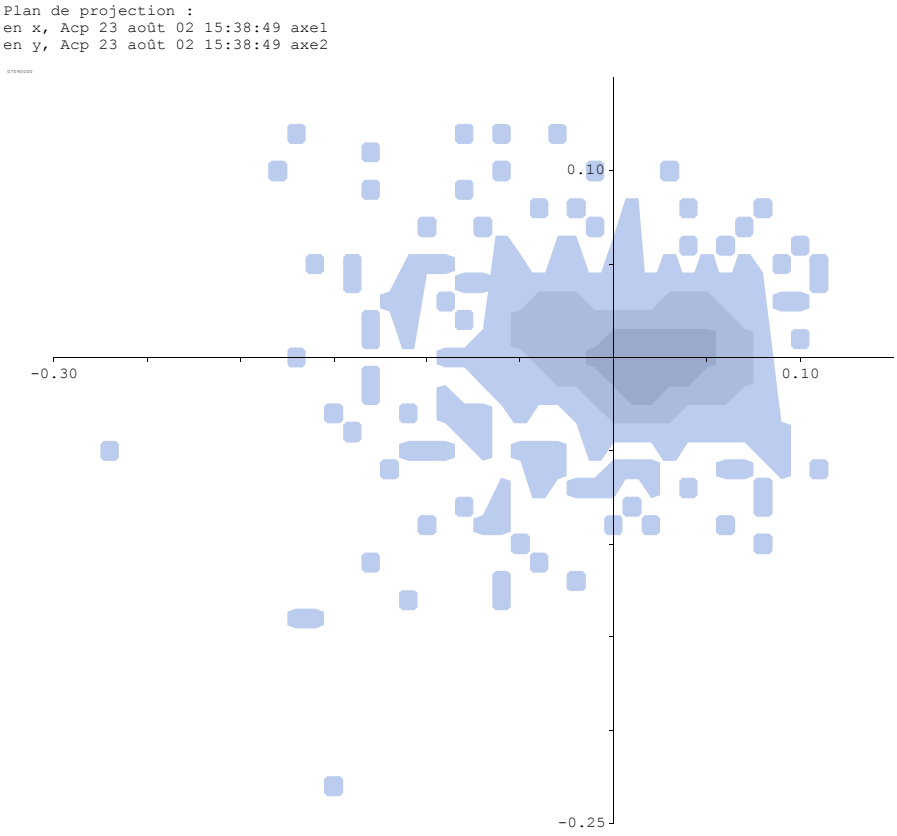
## *Les axes du plan et ce qu'on y projette*

Toute la technique du plan de projection repose sur ceci : si l'on dispose de deux variables numériques définies pour un certain nombre de sujets, alors on dispose pour chaque sujet d'un couple de valeurs (x, y) qu'on peut interpréter comme la position de ce sujet dans un graphe cartésien dont les deux variables constituent les deux dimensions.

## Projection des sujets

On peut dès lors, théoriquement, représenter les sujets par un point ou un signe dessiné à l'intersection de ses coordonnées. Hector n'utilise pas cette possibilité, parce qu'il se destine à traiter des ensembles de données qui peuvent être conséquents, et que dans ce cas rien n'est plus trompeur que cette technique de représentation : si vous voyez un point dans le plan, ça veut dire indifféremment qu'il y a un sujet à cet emplacement, ou qu'il y en a trois cent douze, ce qui peut aisément arriver avec des numériques entières telles que des scores.

De là la nécessité pour Hector de rendre plutôt compte du fait qu'il y a beaucoup de sujets à tel endroit, peu à tel autre et pas du tout ailleurs : la méthode utilisée est, comme dans le croisement de numériques dans le volet [tris], celle du nuage de densité : le « beaucoup » à un endroit étend ses effets autour de lui dans un court rayon, et assombrit la zone. S'il existe d'autres « beaucoup » dans le secteur, ces effets s'accumulent et la zone devient très sombre. En fait, ce que représente une tache, c'est la densité moyenne de sujets dans une zone dont cette tache est le centre.




On peut aussi souhaiter obtenir une représentation différenciée des sujets, selon la valeur qu'ils ont pour une variable du type logique, ordinal ou nominal.

Il faut avoir pour cela installé une variable comme source du « selon » : on la sélectionne dans la liste de variables, et on clique le bouton qui représente un diagramme :

Les différences de valeur peuvent s'exprimer par des couleurs ou par des symboles.

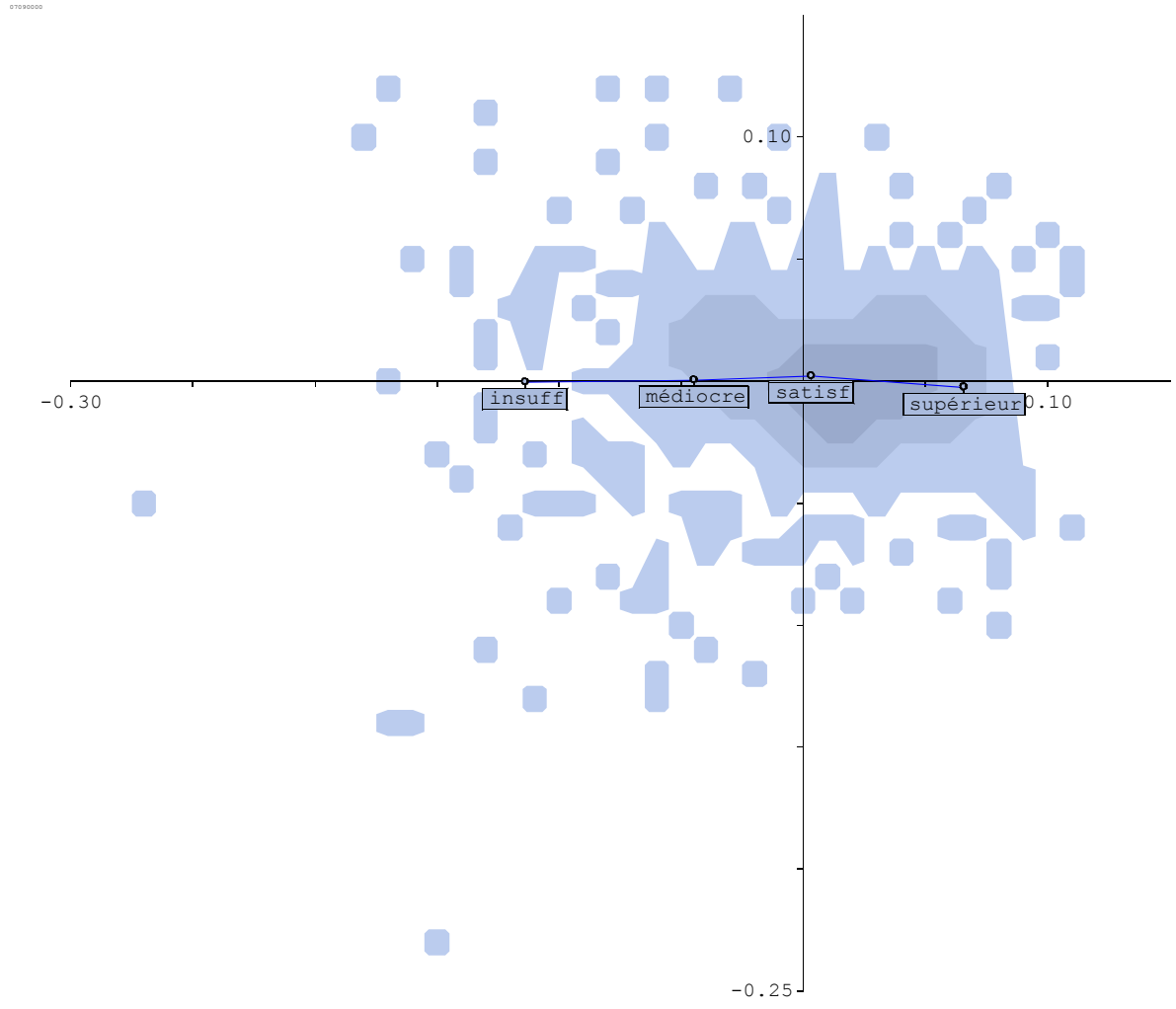
## Projection des variables

*Variables logiques, nominales ou ordinales*

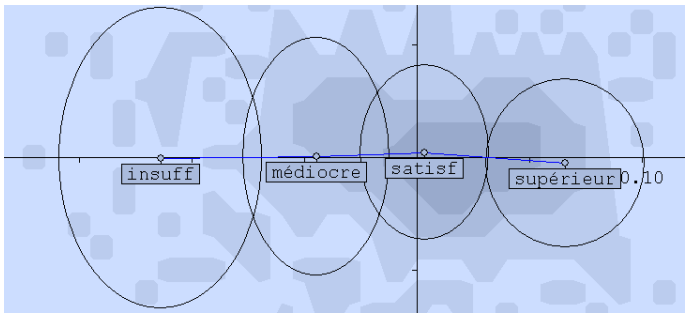
De plus, si on peut connaître la position d'un sujet, on peut calculer la position moyenne d'un groupe de sujets. Les variables qui découpent des catégories dans la population (logiques, ordinales, nominales) fournissent de tels groupes, qu'on pourra représenter en leur position moyenne par l'étiquette de la catégorie. C'est ce qu'on appelle la projection de variables, par opposition à la projection des sujets. On peut projeter autant de variables qu'on veut, la limite étant évidemment dans l'encombrement du schéma : les étiquettes évitent de se recouvrir, grâce à un système de renvoi à des points, mais ce système a lui-même ses limites.

Les deux modes de représentation, densité de sujets et étiquettes de variables, ne sont pas incompatibles et peuvent être utilisées simultanément, comme ci-dessous.

```
Plan de projection :
en x, Acp 23 août 02 15:38:49 axe1
en y, Acp 23 août 02 15:38:49 axe2
```




Ici la variable projetée est ordinale, raison pour laquelle ses positions sont reliées par un trait bleu (c'est une option).




Si la variable est sélectionnée dans la liste, l'affichage s'enrichit, pour chaque position, d'ellipses dont les pseudo-rayons sont proportionnels à l'écart-type de la distribution des sujets relevant de cette étiquette selon les deux variables-axes. Ces pseudo-rayons peuvent être multipliés par un coefficient réglé par le compteur, initialement établi à 100%.

*Variables numériques*

Les variables numériques peuvent également être projetées, il suffit de les ajouter à la liste ào projeter (avec les flèches bleu clair). La représentation d'une numérique consiste en un vecteur dont les coordonnées sont proportionnelles à la corrélation de la variable avec chacun des deux axes.

Le cercle rouge est le cercle unité : une variable dont le vecteur aboutit sur ce cercle est parfaitement représentée dans le plan engendré par les deux axes. Cette projection est particulièrement utile en conjonction avec une Analyse en Composantes Principales.

L'attention du lecteur doit être attirée sur le fait que la technique du plan de projection ne comporte pas d'hypothèse sur l'orthogonalité des variables utilisées comme axes, alors que s'il s'agit de l'utiliser pour exploiter le résultat d'une analyse factorielle, c'est implicitement attendu. C'est donc à l'usager de fournir à son propre lecteur les éléments permettant d'éviter tout malentendu.

## *Mise en œuvre*

Dans tous les cas, il faut désigner les axes, qui ne peuvent être que des variables numériques. On sélectionne une variable numérique dans la liste générale, et on clique sur le bouton [X], puis une autre et on clique sur le bouton [Y]. On peut faire dans l'ordre inverse, mais les deux variables doivent être différentes : si on essaie d'installer en Y la même variable qu'en X, elle s'efface du champ X.

L'usage de la case à cocher [échelles égales] est à considérer en fonction de la nature des variables numériques et surtout de leur ordre de grandeur. Si la case est cochée, le logiciel essaie de représenter l'unité verticale dans la même échelle que l'unité horizontale. Si l'une des variables prend ses valeurs en milliers et l'autre en centièmes, une telle tentative est vouée à l'échec : il vaut mieux utiliser pour chaque variable une échelle indépendante, arbitrairement choisie pour garnir harmonieusement l'espace du schéma.

Ayant désigné les axes, il faut préciser ce qu'on y projette. Si on ne le fait pas, Hector protestera. Les possibilités résultent de la combinaison de la représentation des sujets et de la projection des variables, la seule incompatibilité étant entre le nuage de densité et les options [couleur selon] et [symboles selon].

On sélectionne donc d'abord l'option de représentation des sujets : aucune, densité ou couleurs. Si on a choisi [couleur], il faut indiquer couleur *selon* quoi : on sélectionne une variable et on clique le bouton coloré placé à côté du champ du nom de variable qui gouvernera la couleur. Il s'agit nécessairement d'une variable possédant des étiquettes implicites ou explicites, et donc de

type logique, ordinal ou nominal. Les couleurs sont gouvernées par le panneau des options : il peut donc s'agir aussi de grisés ou de camaïeux.

On sélectionne ensuite éventuellement des variables à projeter. Il s'agit aussi de variables logiques, ordinales ou nominales : on les sélectionne dans la liste générale, et on clique le bouton à la flèche entrante. Le nom de la variable vient s'inscrire dans la liste des variables à projeter. Inversement on peut sélectionner une variable dans cette liste, et l'enlever d'un clic sur le bouton à la flèche sortante ; on peut aussi vider la liste en cliquant la poubelle.

Il n'est pas obligatoire de projeter des variables, sauf dans le cas où on a choisi de n'utiliser aucune représentation des sujets, puisque alors il n'y aurait rien à dessiner. En revanche, la projection de variables est compatible avec toutes les options de représentation des sujets. La variable éventuellement utilisée pour [couleur selon] peut aussi être projetée, mais ce n'est pas une obligation.

On peut projeter autant de variables que l'on veut : théoriquement, pas plus de cent, mais la vraie limite intervient bien avant, à cause de l'illisibilité du schéma.

# EXPLOITATION DES RESULTATS

Les résultats tabulaires, statistiques et graphiques produits par Hector sont lisibles à l'écran, imprimables directement si on le souhaite, mais le mode historique de transfert et de récupération des résultats est l'usage du document. On verra ci-après que d'autres procédés sont disponibles.

## *La notion de document*

Indépendamment du format de sortie choisi, certains principes gouvernent l'usage du document.

Ce qu'on appelle ici Le Document est concrètement, un fichier, dont le format peut varier selon le système d'exploitation, mais correspond toujours à des spécifications publiées. Il est destiné à être repris dans un logiciel de traitement de texte pour y constituer un élément d'un ensemble plus vaste (rapport, article, thèse…). Les éléments de texte, de tableaux et de graphiques sont *vivants*, c'est-à-dire qu'ils peuvent être réédités, enrichis, agrémentés, mais aussi modifiés : Hector n'est pas responsable de ce qu'on fait ensuite de ses productions. Pour ce faire, les images, notamment, sont en mode vectoriel, c'est-à-dire exprimés par une description géométrique analytique indépendante de la taille finale, et non pas en mode pixel, comme le serait une photographie.

L'usage de deux majuscules dans l'expression Le Document n'est pas lié à un goût particulier pour l'emphase, mais au fait fondamental qu'il n'y a jamais plus d'un document à la fois. Plusieurs fichiers issus de l'envoi au document peuvent bien sûr coexister, mais un seul, à un moment donné, est le document actif.

Au lancement d'Hector, le bouton d'accès au document, dans la page TRAITEMENTS (et quel que soit le volet actif), porte une flèche accompagnée de la mention [document à définir] : aucun document n'est encore défini. Cela n'a pas d'importance tant qu'on ne souhaite pas y envoyer quelque chose.

Quand, ayant des résultats à envoyer, on clique ce bouton, il donne accès à un dialogue d'ouverture de fichier, à l'issue duquel le nom du fichier désigné garnira le bouton. Si le fichier désigné existait déjà, un dialogue demande confirmation de ce choix, parce que la première chose que fait Hector est d'en effacer le contenu antérieur.

Ensuite, tant qu'on clique ce bouton, on envoie à la suite du document ce qui est affiché à l'écran. Cela ne cessera qu'à la fermeture du document.

Le document se ferme automatiquement quand on quitte Hector, mais on peut aussi le fermer explicitement en droite-cliquant le bouton, qui reprend la mention [document à définir].

Quel intérêt peut-il y avoir à fermer un document en cours de session Hector ?

- Une première raison est du type « Ne pas mettre tous ses œufs dans le même panier ». Autrement dit, une session peut être longue et produire beaucoup de sortie, et donc un document très long. On peut se retrouver dans une situation analogue à celle des années 70, où l'on apportait son paquet de cartes perforées codant les demandes de traitements à Monsieur l'Informaticien, qui vous rendait après quelques heures ou quelques jours selon votre rang social les douze kilos de listing correspondant à votre requête, paperasse qu'il vous restait à dépouiller pour vous apercevoir que vous vous étiez fourvoyé dans la méthode dès la quatrième ligne, et que tout était à refaire. Ce genre de joyeuseté à déterminé la vocation de l'auteur pour le logiciel statistique interactif. Dans cet esprit, le document, non content d'être un moyen de transport, peut constituer un second lieu de tri et de réorganisation ou de stockage intermédiaire. La

production du rapport final est rarement un processus linéaire : on n'expose pas nécessairement les résultats dans l'ordre où on les a constitués. Un premier tri s'exerce devant l'écran : on envoie ou on n'envoie pas au document ; le second tri est qu'on peut envoyer les éléments liés à un aspect donné de la question dans un document, puis le fermer et en ouvrir un autre pour traiter un autre aspect ; le troisième tri intervient au moment de l'insertion des documents dans l'ouvrage, de l'agencement des parties, etc.… Cette méthode de travail peut s'avérer étonnamment souple, à condition de savoir nommer efficacement ses fichiers-documents, de les répartir judicieusement dans des dossiers et sous-dossiers, et sans doute le mieux, d'en tenir trace dans un journal de bord du travail de recherche dont on ne saurait trop conseiller la tenue.

- Le second intérêt est plus technique. Un fichier document qui vient d'être produit est immédiatement disponible pour consultation avec un traitement de texte, même en laissant Hector ouvert, à la seule condition que ce document soit fermé, parce que peu de systèmes informatiques admettent la mainmise simultanée de deux logiciels (Hector et le traitement de texte) sur le même fichier (le document). Cela peut s'avérer utile quand on veut vérifier rapidement les effets sur le rendu final de divers réglages et paramètres concernant les tableaux et les graphiques.

Le fichier document peut être la plupart du temps ouvert directement avec un traitement de texte, mais sa destination est plutôt d'être inséré (commande genre Insérer Fichier …) dans un texte déjà élaboré. En effet, le document ne comporte pas d'indications de taille de papier, d'orientation de l'imprimante, etc., et si le traitement de texte n'est pas suffisamment souple, il risque de refuser d'imprimer tant qu'on ne lui fournit pas du papier au format US Legal au lieu de l'européen A4 (en tous cas ça s'est produit dans le passé).

---

## *Document au format .rtf*

Le choix du format de document s'opère dans le panneau de configuration et d'habitudes, accessible depuis la page d'accueil CORPUS par le bouton [Options de configuration et Habitudes], dans le cadre [Généralités].

Ce panneau comporte un cadre [Paramètres du document de sortie], lequel comporte entre autres un sélecteur [Format du fichier], qui offre trois possibilités : RTF, ODT, TEX et HTML.

Le nom du format RTF signifie Rich Text File : il s'agit d'une convention d'enrichissement des textes issue du monde de l'imprimerie. Cette convention est largement répandue dans différents systèmes, y compris le Mac.

Malheureusement, on ne dispose pas d'une convention équivalente pour les graphiques vectoriels, et on s'est trouvé contraint d'utiliser ici, pour les images, le format WMF ou EMF, qui est propre à l'univers Windows. De ce fait, Hector ne peut utiliser le format RTF que dans cet univers.

Le format ODT est le format de la série bureautique libre Open Office.

Le format TEX est en principe utilisable sous divers systèmes d'exploitation.

Le format HTML (HyperText Markup Language) permet de publier directement des résultats d'Hector sur Internet.

---

## Sorties sans mise en forme

L'une des options du cadre [Paramètres du document de sortie] est la case à cocher [Tableaux mis en forme]. Si elle n'est pas cochée, la sortie est –relativement – brute.

(OPTION)

| | effectifs | %/Total |
|---|---|---|
| com educ | 129 | 13,34% |
| didact | 373 | 38,57% |
| docum | 36 | 3,72% |
| inconnue | 62 | 6,41% |
| w sansoc | 367 | 37,95% |
| Total | 967 | 100.00% |

Efficacité entropique : 80,9%

Un ensemble de textes, tableaux et statistiques, qui, à l'écran, avaient l'aspect suivant :

se retrouvera, par exemple dans le traitement de texte Word, sous cette forme :

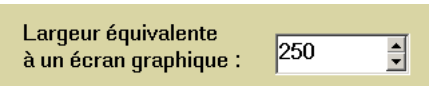
(OPTION)¶
→ ·effectifs···%/Total¶
··com·educ → ·····129→····13,34%¶
····didact·····373→····38,57%¶
·····docum → ·····36→·····3,72%¶
··inconnue → ·····62→·····6,41%¶
··w·sansoc → ·····367→····37,95%¶
·····Total·····967→···100.00%¶
Efficacité·entropique·:·80,9%¶
¶

On a utilisé l'option de Word qui rend visibles les caractères de mise en page. Les colonnes du tableau sont séparées par des tabulations (matérialisées ici par des flèches), et la police de caractères est la police par défaut du logiciel, vraisemblablement une sorte de Times Roman. Les petits points à mi-hauteur représentent des espaces.

L'idée qui sous-tend une telle sortie brute des textes et tableaux est qu'il est extrêmement aisé à un utilisateur maîtrisant correctement son traitement de texte d'obtenir à partir de là une mise en forme du tableau élégante et de son choix parmi une multitude de possibilités de style.

En revanche les éléments graphiques sont d'ores et déjà codés en wmf/emf, insérés dans le texte rtf.

Une des options proposées intéresse directement les graphiques :




Cette largeur équivalente est réglable, et permet notamment d'autoriser de s'adapter si on dispose d'un écran très grand avec une définition très fine : des graphiques qui dans une autre configuration ne tiendraient pas en entier dans l'écran (et feraient apparaître un ascenseur latéral) y sont très à l'aise, mais il faut « prévenir » le traitement de texte que les images sont larges en taille pixel.

Du reste, il ne s'agit que de l'aspect initial d'une image lors de l'importation dans un traitement de texte, puisque les images, vivantes, y sont facilement réajustables en taille.

## Sortie avec mise en forme

si l'option de mise en forme est cochée, la sortie du même tableau a plutôt l'aspect suivant :

(OPTION)

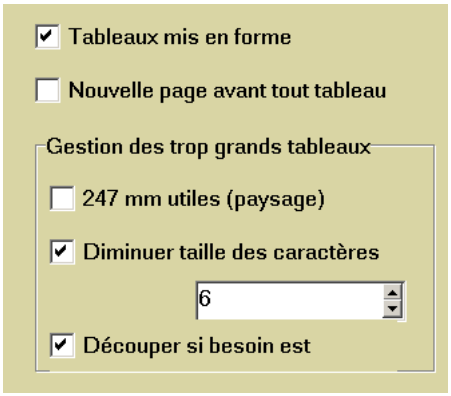
| | effectifs | %/Total |
|---|---|---|
| com educ | 129 | 13,34% |
| didact | 373 | 38,57% |
| docum | 36 | 3,72% |
| inconnue | 62 | 6,41% |
| w sansoc | 367 | 37,95% |
| Total | 967 | 100.00% |

Efficacité entropique : 80,9%

Différences essentielles : une police de caractères à chasse fixe (Courier New en 10 points) est imposée aux textes, et, sauf aménagement, aux tableaux, et ces derniers sont dessinés avec leurs cadres et leur alignement.

Cette mise en forme des tableaux est la plus rustique qu'on puisse rêver, mais elle dispense – temporairement- de se préoccuper de cet aspect des choses.

De plus, quand l'option est active, elle donne accès à d'autres options, dont un mécanisme de gestion des grands tableaux :




La seconde case à cocher propose d'insérer un saut de page entre deux tris ou croisements.

La gestion des grands tableaux, ceux qui sont trop larges pour tenir dans la page, dispose de deux ressources principales : la diminution de la taille des caractères et le découpage des tableaux en tranches successives.

La case à cocher [247 mm utiles (paysage)] fournit un point de repère à ces deux ressources : sommes-nous dans une page A4 en portrait : 160 mm utiles avec les marges usuelles de 25 mm, ou en A4 paysage : 247 mm utiles avec les mêmes marges ?

Si un tableau menace d'être trop grand, et que la case [Diminuer taille des caractères] est cochée, Hector cherche d'abord si, en diminuant la taille des caractères (s'entend du haut des majuscules au haut de celles de la ligne suivante en interligne simple) par demi-points à partir du standard initial de 10 points (il y a 72 points typographiques dans un pouce, qui vaut 25,4 mm, d'où une valeur du point d'environ 0,35 mm) et jusqu'au minimum indiqué juste en dessous, il serait possible de faire tenir le tableau dans une ligne.

Si la diminution des tailles de caractères n'est pas autorisée, ou si elle ne suffit pas, et que la case [Découper si besoin est] est cochée, intervient le découpage.

Hector compte en combien de tranches il lui faudra découper le tableau, et remonte l'échelle des tailles de caractères, sans dépasser 10 points ni augmenter le nombre de tranches.

(SEXE) × (PROFESSION)

| N | anim sc | ass soc | aucune | autres | cadreFI | cadreHE | educ spe | form ad |
|---|---|---|---|---|---|---|---|---|
| femmes | 65 | 14 | 255 | 36 | 14 | 9 | 18 | 6 |
| hommes | 11 | 3 | 61 | 8 | 2 | 3 | 20 | 2 |
| S/COLONNE: | 76 | 17 | 316 | 44 | 16 | 12 | 38 | 8 |

| inconnue | insti | m.a | mait.for | mi/se | prof | sante | S/LIGNE : |
|---|---|---|---|---|---|---|---|
| 173 | 32 | 7 | 6 | 61 | 18 | 24 | 738 |
| 56 | 12 | 1 | 3 | 31 | 13 | 2 | 228 |
| 229 | 44 | 8 | 9 | 92 | 31 | 26 | 966 |

Le tableau ci-dessus, qui comporte 17 colonnes, a été découpé en deux tranches, avec une taille de caractères réduite à 7,5.

Le réglage du minimum de taille de caractères est par défaut à 6, mais peut descendre à 5. Ce sont alors de très petits caractères, qui n'ont de sens que si on a l'intention de procéder à des agrandissements de l'impression A4.

Si la réduction de taille n'est pas autorisée ou ne suffit pas et que le découpage n'est pas autorisée, le tableau excède la largeur de la page A4 portrait ou paysage et voilà tout. L'usager peut encore le découper à la main, ou peut-être dispose-t-il d'une imprimante A3 ou plus.

## *Document au format .odt*

L'auteur a implémenté ce format de sortie en réponse à la demande de nombreux usagers, notamment les étudiants impécunieux qui ne possédaient pas de licence pour Word. Utiliser le logiciel libre n'est nullement incompatible avec la philosophie de l'auteur, mais certaines difficultés, notamment de documentation, font qu'en dépit d'efforts constants il peut arriver que la sortie .odt ne soit pas tout à fait aussi jolie que la sortie .rtf, qui est le format initial.

Le gros avantage du format .odt, outre la gratuité d'Open Office, est évidemment que le Document est produit dans un format officiel et public. Pour les curieux, on notera que le format .odt est en fait la compression d'un système de fichiers au format .xml, lesquels peuvent petree lus comme du texte. Pour s'en assurer, il suffit de changer l'extension du fichier de .odt en .zip, en ignorant les protestations du système, puis de dézipper pour découvrir les fichiers qui le composent.

## *Document au format .tex*

> L'auteur est redevable de la rédaction de cette rubrique à Lionel Conraux, Maître de Conférences en Sciences de l'Education et *très-sçavant* en choses de la mise en forme des documents en général et de $T_E\chi$ en particulier.

Par défaut, Hector est configuré pour produire des sorties au format RTF. Il est possible d'obtenir des fichiers sources compatibles avec LaTeX en activant la case [tex] dans la zone correspondant au réglage des paramètres du Document de sortie, accessible à partir de l'item [Options de configuration et Habitudes] de la page CORPUS. Ce paramétrage du logiciel peut être rendu permanent en enregistrant les options de traitement.

Les tests ont été effectués sous Windows®, avec la distribution TexLive, version 7 diffusée par l'association GUTenberg des utilisateurs français de TeX.

Le traitement de texte LaTeX (Lamport, 1985) impose la définition d'un préambule lors de la production de tout document valide. Le ou les documents de travail produits par Hector sont des documents complets, *compilables* : ils commencent donc par un en-tête, contiennent les instructions nécessaires à la réalisation des sorties demandées dans un bloc débutant par

`\begin{document}` et se terminent par `\end{document}`. L'architecture générale d'une sortie Hector est la suivante : le fichier commence par un commentaire indiquant le producteur du document, il est suivi par l'en-tête ; quelques indications générales suivent la déclaration du début du document. Les sorties demandées figurent après la commentaire `Corps du document`.

```
% Sortie automatique Hector -- AD
[en tête du document]
[…]
\begin{document}
[…]
% Corps du document
statistiques globales, tableau et/ou graphique
\end{document}
```

## En-tête du document

L'en-tête du document fixe la classe du document produit (article), le format de papier (A4), la taille du corps de la police utilisée (10 points). Afin de minimiser les débordements de la feuille de papier, on impose une réduction des décalages horizontal et vertical de un pouce. On déclare ensuite l'extension utilisée lors de la production des graphiques : `pstricks`. Le document produit comporte du texte rédigé en français, il est adapté à cette langue (package `Babel` (Johannes Braams , 2001, version 3.7) avec l'option `francais`) et on utilise le codage `T1` permettant d'utiliser en 8 bits les caractères accentués nécessaires. Ces options fixées lors du développement produisent l'en tête suivant :

```
% Sortie automatique Hector -- AD
\documentclass[a4paper,10pt]{article}
\usepackage{pstricks}
% En tête du document
\usepackage[T1]{fontenc} % caractères accentués au clavier
\usepackage[francais]{babel} % tout en français
\setlength{\hoffset}{-1in}
\setlength{\voffset}{-1in}
```

Puisque le document produit par Hector est un texte pur compilable par LaTeX, il est possible de modifier ce paramétrage, en assumant les conséquences des modifications effectuées : ainsi, augmenter la taille de la police utilisée, réduire le rectangle d'empagement pourraient conduire à des débordements et/ou à des positions inadaptées d'étiquettes dans les graphiques. Le concepteur d'Hector ne saurait être tenu pour responsable des aléas résultants de telles modifications…

Après l'en tête figure le corps (au sens LaTeX) du document. On commence par sélectionner une police de taille fixe (`\ttfamily`) et indiquer à l'extension `Babel` que la langue utilisée est le français. Ces quelques indications se terminent par un commentaire indiquant le début réel du corps du document. Cette section du source du document correspond au texte source suivant.

```
\begin{document}
\ttfamily
\selectlanguage{francais}
\frenchspacing
% Corps du document
```

Les sorties produites par Hector peuvent être des statistiques globales (moyenne, mode, médiane, écart-type, …), un tableau (simple ou croisé) ou un graphique (histogramme, camenbert, …).

## Statistiques globales

Les statistiques globales relatives à une variable donnée sont produites sous la forme d'un alinéa se terminant par la commande `\par`.

```
Efficacité$\;$entropique$\;$:$\;$83,6\%\par
```

Les " texniciens " reprocheront certainement l'usage abusif d'espaces en mode mathématique, trace visible d'un ancêtre produisant des sorties au format RTF. Ce défaut de jeunesse systématique qui affecte toutes les productions pourrait disparaître dans une version ultérieure.

## Tableaux

Afin de permettre une portabilité maximale et des sorties aussi semblables que possibles en TeX (Knuth, 1986) et en rtf, les tableaux produits n'utilisent aucun décor particulier : ni l'extension `Booktabs` (Simon Fear, 1995) produisant des tableaux d'une qualité typographique supérieure, ni l'extension `colortab` (Timothy Van Zandt , 1997) permettant l'usage de la couleur ne sont, par exemple, employéees. On utilise l'environnement `tabular` : chaque tableau produit est précédé par une indication textuelle indiquant la ou les variables concernées. De manière systématique, chaque case d'un tableau est délimitée par des filets horizontalement et verticalement ; les données sont alignées à droite. Le source type d'un tableau est conforme au schéma suivant :

```
(Domaine$\;$diplôme$\;$antérieur)\par
\begin {tabular} {| r | r | r |}
\hline
première ligne \\
…
dernière ligne du tableau \\
\hline
\end{tabular}
```

Hector ne gère pas un éventuel débordement du rectangle d'empagement d'un tableau qui comporterait de nombreuses colonnes et/ou lignes. L'utilisation de l'extension `Longtable` (David Carlisle, 1990-1998) est laissée à l'initiative du producteur du document final. Il en est de même pour une éventuelle rotation d'un tableau.

## Graphiques

Les graphiques générés par Hector utilisent l'extension `pstricks` (Timothy Van Zandt, 1993). Chacun d'entre eux est contenu dans un environnement `pspicture` indiquant la taille horizontale et verticale (en centimètres) du dessin. Chaque graphique exporté est donc conforme au schéma suivant~ :

```
\begin{pspicture}(6.20,6.79)
[instructions de dessin vectoriel …]
\end{pspicture}
```

La mise à l'échelle peut s'effectuer sans modifier les coordonnées des points en insérant une instruction inspirée de celle indiquée ci-dessous qui réduit à 75% le graphique produit. Dans ce cas, il est nécessaire de modifier la taille déclarée du schéma ou de la faire calculer par LaTeX en utilisant les services d'une extension comme Realcalc (Frank Buchholz, Jan. 1993) ou Calc (Kresten Krab Thorup & Frank Jensen, 1992—1995).

```
\begin{pspicture}(4.65,5.09) % 6.20*0.75,6.79*.75
\psset{xunit=.75,yunit=.75}
[instructions de dessin vectoriel …]
```

```
\end{pspicture}
```

On notera, en particulier, qu'un graphique obtenu à l'issue d'une analyse de séquences (Dubus 2000) utilise une police en chasse fixe de petite taille.

## *Document au format .html*

La sortie au format HTML présente plusieurs particularités :

- les hachures n'y sont pas utilisées, mais, s'agissant de documents préférentiellement destinés à la lecture à l'écran, les couleurs sont conseillées.
- les graphiques ne sont pas dans le fichiers HTML, mais dans des fichiers SVG (Scalable Vectors Graphics) qui l'accompagnent, à raison d'un fichier SVG par graphique. Ces fichiers ont le même nom que le fichier HTML, avec des suffixes _1, _2 …
- les fichiers SVG satellites d'un fichier HTML sont réputés placés dans le même répertoire (chemin) que le HTML. Ne pas l'oublier en cas de mise en ligne ou d'envoi par courrier électronique
- les fichiers HTML et SVG sont rédigés en texte, dans une syntaxe publique et disponible partout sur la toile. On peut donc parfaitement les modifier à la main, avec un éditeur texte, à condition de savoir ce qu'on fait. On trouve également des éditeurs visuels de HTML et de SVG, certains gratuits, d'autres non. Il est donc possible de retoucher et de mettre en page textes et images.
- pour visualiser correctement les images SVG dans les fichiers HTML, il faut disposer d'un logiciel « plug in » SVGviewer, disponible gratuitement sur le site d'ADOBE. Ce logiciel s'acoquine tout seul avec le navigateur utilisé, pour rendre visibles les images SVG. Ne pas oublier de le faire savoir aux gens à qui on envoie de telles images.

Les sorties HTML sont malheureusement sensibles à l'évolution des navigateurs et à la version du plug-in utilisé, et en peuvent donc être garanties en l'état.

## *L'impression directe*

Le bouton [vers l'imprimante] des pages TRAITEMENTS et FACTORIELLES permet d'envoyer à l'imprimante ce qui est actuellement affiché comme résultat.

L'imprimante utilisée, si on en possède plusieurs, est celle qui est actuellement définie comme imprimante par défaut dans le système d'exploitation. Hector ne propose pas d'organe pour en changer, mais il est aisé de le faire en parallèle.

En revanche Hector propose, dans le panneau des Options de configuration et habitudes, accessible depuis la page CORPUS, la possibilité de fixer les marges de l'imprimante (15 mm par défaut).

L'impression directe est intéressante quand on a un besoin immédiat des résultats, par exemple dans un travail collectif. L'utilisateur préférera souvent passer par le document, qui lui offre des possibilités plus fines de réglage des tailles, notamment en ce qui concerne les graphiques.

## *Le copier-coller*

A côté du procédé ancien mais robuste du document, Hector disposait jusqu'à une époque récente d'un copier-coller direct depuis une page de résultats vers un traitement de texte, en droite-cliquant dans l'image de résultat, puis en collant (CTRL-V) dans le traitement de texte ouvert. Ce procédé, rapide, présente cependant un inconvénient : ce qui est transporté est une

image globale insécable, sauf au prix de pénibles contorsions avec plusieurs logiciels : il est impossible de la découper, de la répartir sur plusieurs pages, et plus encore de la modifier, alors que le procédé du document de transport fournit des textes, tableaux et graphiques « vivants », c'est-à-dire modifiables à souhait.

Les nouveaux mécanismes de copier-coller présentés ici ne remplacent pas totalement le document, qui reste la seule bonne solution dans certains cas, mais proposent une alternative rapide et efficace dans la plupart des situations courantes.

## Les variantes du copier-coller

On désigne ici par copier-coller le fait de cliquer (avec l'un ou l'autre bouton et l'une ou l'autre touche de contrôle) dans une page de résultats et de récupérer dans le « presse-papiers de Windows », ce réceptacle invisible du transport de morceaux de documents tout ou partie du résultat pour le coller dans un autre logiciel de traitement de texte ou autres. Le procédé a été testé avec Word, Excel, Power Point, Corel Draw, et les éléments d'OpenOffice.org : Write, Calc, Draw, Impress. Il est probable qu'il fonctionne avec tout logiciel de ce genre sous Windows, puisqu'il fait usage de fonctionnalités standard.

Le but de la manœuvre est de travailler en parallèle avec deux logiciels ouverts simultanément, Hector sur lequel on traite des données, et un autre où l'on compose au fur et à mesure le commentaire.

Selon les logiciels et leurs versions, *il peut être nécessaire que le logiciel d'accueil ait été ouvert avant* que l'on commence à copier des éléments. C'est donc une précaution utile dans tous les cas.

Quatre variantes sont possibles :

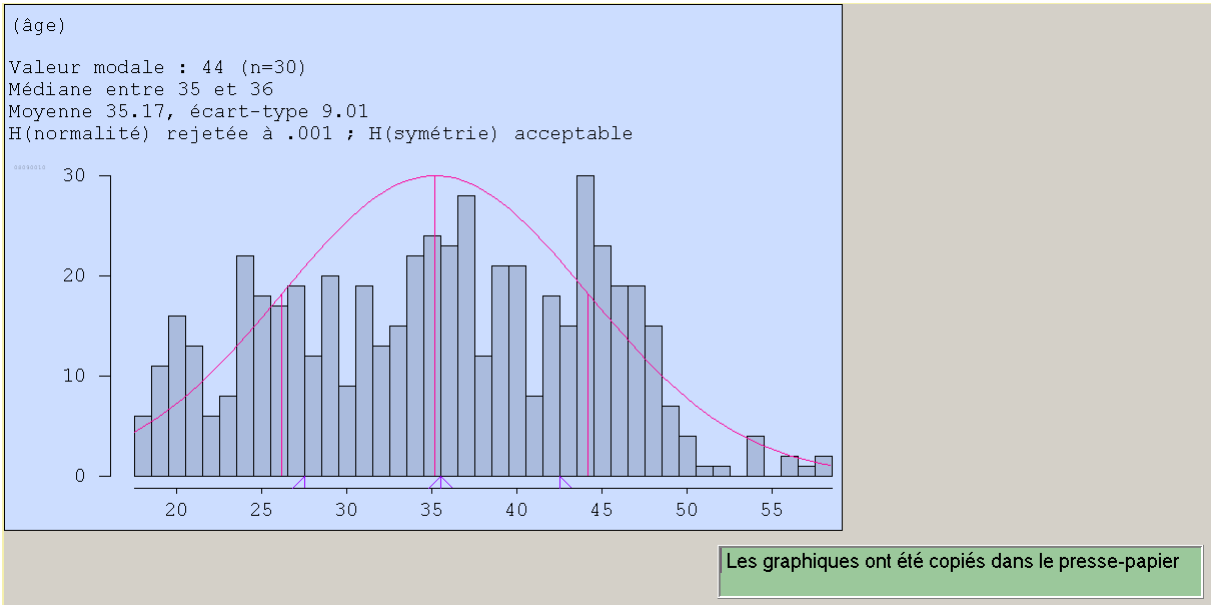
- Avec un clic-droit, c'est le copier-coller déjà connu, qui emporte tout le résultat, graphiques, tableau et textes groupés ensemble dans une seule image insécable. On n'en reparlera pas, puisque les nouveaux dispositifs visent précisément à remédier à ses défauts.
- Avec un clic-gauche simple (sans touche de contrôle), seuls les graphiques sont copiés dans le presse-papiers.
- Avec un Ctrl-clic-gauche, seuls les textes et les tableaux sont copiés dans le presse-papiers
- Avec un Alt-clic-gauche, variante du précédent, seuls les textes et tableaux sont copiés, mais avec un formatage spécial pour certains types de textes qui sont aussi un peu des tableaux.

Le but de cette décomposition des types d'objets transportés est de permettre leur modification, mais aussi et surtout de les séparer pour pouvoir y entremêler les commentaires.

## Copier-coller une image seule

On gauche-clique l'image du résultat, et un message informe que les graphiques ont été collés dans le presse-papier. Il ne peut y avoir qu'un seul objet (image ou série de textes) à la fois dans ce genre de presse-papiers.

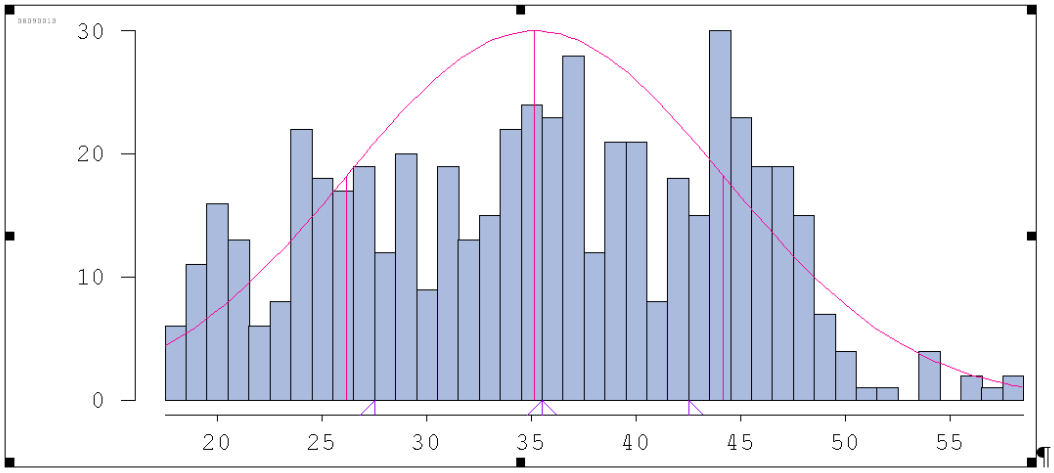
Par exemple, on trié (sans les résultats tabulaires) une variable « âge », et quand on clique dans l'image du résultat avec le bouton gauche de la souris :




Le message d'Hector s'affiche brièvement pour signaler que la copie a été effectuée.

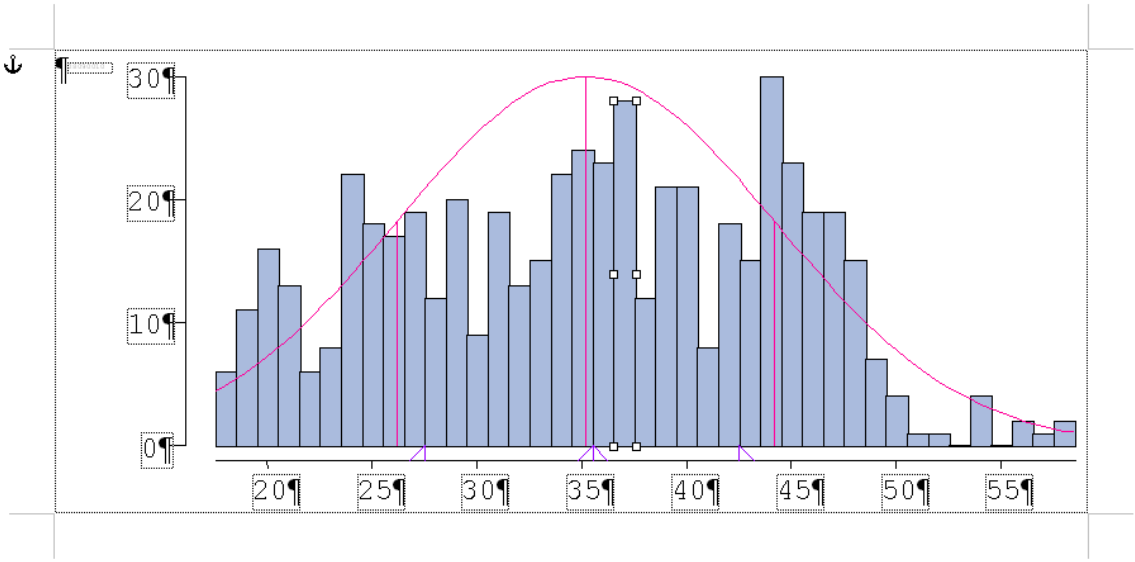
*Chez Microsoft*

Si on vient maintenant cliquer dans Word © par exemple, et qu'on colle avec Ctrl-V, on obtient ceci :

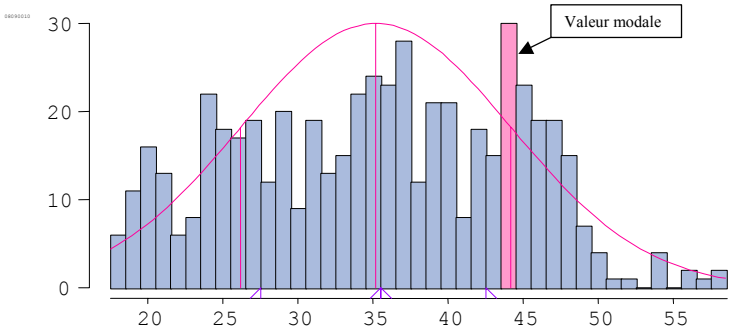



Les carrés noirs, ou poignées de contrôle, sont visibles, ainsi que le cadre, parce qu'on a cliqué dans l'image copiée, pour la sélectionner. Ils ne sont pas visibles le reste du temps.

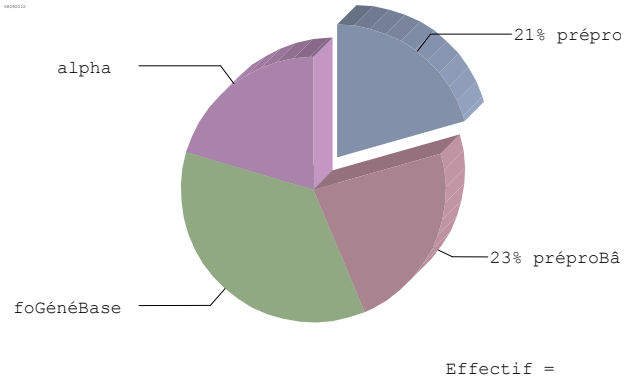
On peut changer la taille ou la forme de l'image en tirant sur ces carrés noirs. Sous Word ©, les carrés de coin respectent les proportions d'origine. On peut aussi cliquer avec le bouton de droite de la souris dans l'image, et sélectionner dans le menu « Format de l'image » pour accéder à des réglages plus précis. On peut aussi sélectionner « Modifier l'image », et on accède alors à l'éditeur graphique (ça marche aussi en double-cliquant l'image) :




On peut y apporter les modifications de détail souhaitées, par exemple pour attirer l'attention sur un détail. Quand c'est fini, on ferme la fenêtre de dessin.

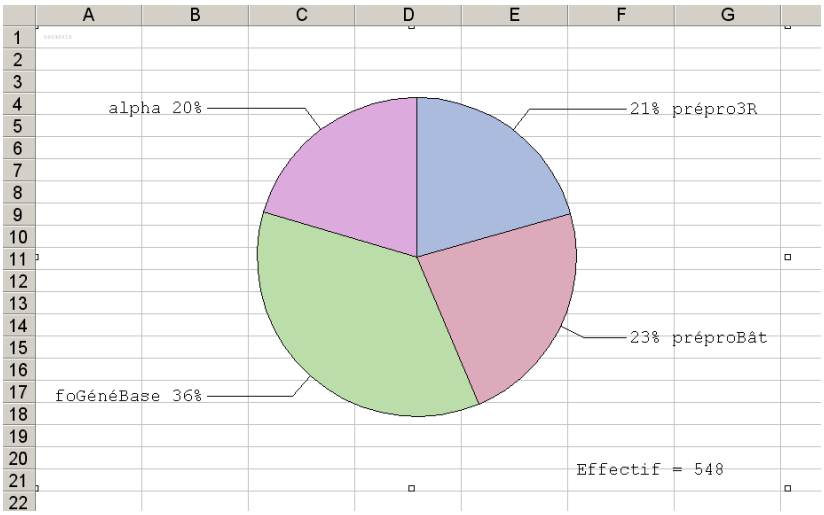



Ici, on a peint en rose et commenté la valeur modale de la distribution. L'usager expérimenté peut sûrement faire beaucoup mieux. Attention cependant : dans l'image en édition, les éléments (un rectangle, un morceau de texte) sont mutuellement indépendants, et si on en déplace un, il peut être très difficile de retrouver sa place. Là, on a 3D-ifié et éclaté le camembert. Y en a qui aiment.




Dans Word ©, l'image ainsi collée est considérée comme un caractère et suit automatiquement le texte. Toutefois cela peut dépendre de la version dont on dispose, et on ne peut exclure qu'elle apparaisse comme une image ancrée. Il est un peu difficile ici de prévoir toutes les variantes des logiciels et leurs évolutions.

Dans Excel ©, cela marche tout aussi bien :




à condition toutefois qu'Excel © ait été ouvert avant le copier-coller (en tous cas avec la version dont on dispose ici et maintenant). Il ne semble pas qu'il soit possible de modifier l'image sur place, mais de nombreuses options affectent son cadrage et son ancrage aux cellules.

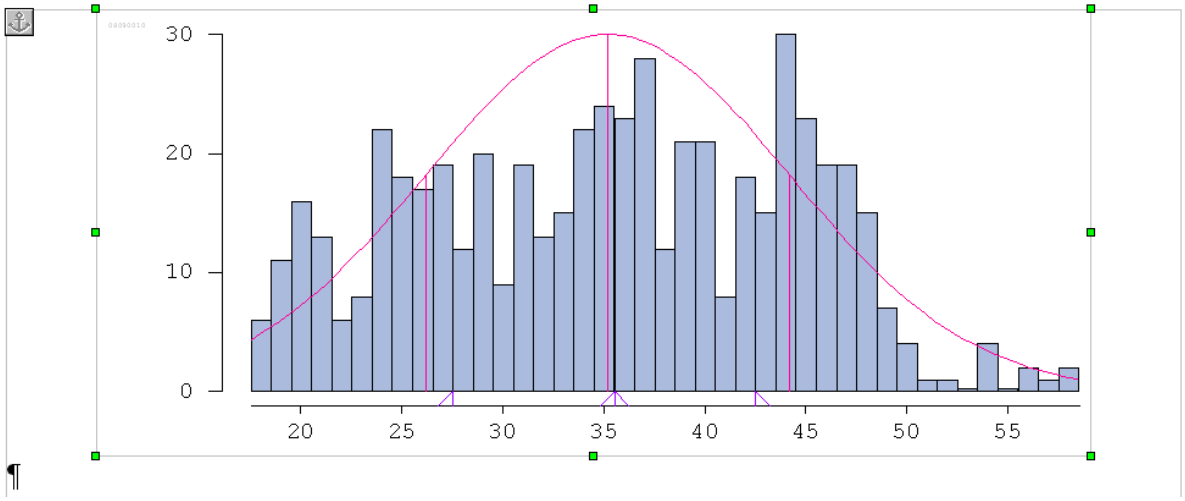
Cela marche aussi avec PowerPoint ©, et là on peut éditer l'image.

*Plusieurs images*

Quand le résultat comportait plusieurs graphiques séparés (cas de l'ANOVA par exemple), les graphiques sont envoyés en une fois comme une seule image. Il est possible de les séparer par l'édition graphique, mais cela est un peu délicat, et le copier-coller est un instrument pour gens pressés. On préférera donc souvent jouer des options de tri pour n'envoyer qu'un graphique à la fois, ou encore recourir à la solution classique du document.

*Chez OpenOffice.org*

Dans l'OpenOffice Writer, on obtient ceci :




C'est à peu près la même chose que dans Word ©, mais l'image est explicitement ancrée, elle est centrée, et, attention ! Pour se servir des poignées (vertes ici) sans déformer, il faut tenir la touche majuscule enfoncée pendant qu'on tire sur les poignées.

La modalité d'ancrage, et bien d'autres choses, peuvent être modifiées en cliquant à droite ou en double-cliquant l'image. Il semble en revanche impossible de modifier l'image dans Writer même. Il faut plutôt passer par l'ouverture d'un document de dessin (OpenOffice Draw) où l'on colle l'image pour la modifier à volonté, avant de la copier sous Draw pour la coller sous Writer. Pour les modifications sous Draw, il peut être utile de Fractionner l'image (Menu Modifier) avant de gérer les détails, et de Grouper l'ensemble avant de copier pour coller sous Draw.

Outre Writer et Draw, on peut aussi coller l'image dans Calc (sans possibilités d'édition, mais c'est la philosophie de OpenOffice de spécialiser le dessin vectoriel dans Draw, plutôt que de le rendre accessible partout), et dans Impress, l'outil de présentation.

## Coller textes et tableaux

*Avec Ctrl-clic-gauche, copie normale*

On a trié une variable, ici sans affichage graphique (mais s'il y avait des graphiques, ils seraient ignorés) :

(filières formation)

| | effectifs | %/Total |
|---|---|---|
| prépro3R | 113 | 20.62% |
| préproBât | 126 | 22.99% |
| foGénéBase | 197 | 35.95% |
| alpha | 112 | 20.44% |
| Total | 548 | 100.00% |

Efficacité entropique : 97.8%

Avec un Ctrl-clic-gauche dans la zone de résultats, on obtient brièvement le message « Les textes et tableaux ont été copiés dans le Presse-papiers ». Collé dans Word ©, on obtient ceci :

(filières·formation)¶
→ ·effectifs → ···%/Total¶
··prépro3R → ·····113 → ····20.62%¶
·préproBât → ·····126 → ····22.99%¶
foGénéBase → ·····197 → ····35.95%¶
·····alpha → ·····112 → ····20.44%¶
·····Total → ·····548 → ···100.00%¶
Efficacité·entropique··:·97.8%¶

On a utilisé l'option qui montre les marques de mise en page et copié l'écran. Les signes de paragraphe sont des sauts de ligne , les flèches des tabulations et les points à mi-hauteur de l'œil des minuscules, des espaces. Sans cette option, on a l'aspect suivant :

(filières formation)

| | effectifs | %/Total |
|---|---|---|
| prépro3R | 113 | 20.62% |
| préproBât | 126 | 22.99% |
| foGénéBase | 197 | 35.95% |
| alpha | 112 | 20.44% |
| Total | 548 | 100.00% |

Efficacité entropique : 97.8%

En d'autres termes, les textes apparaissent comme des textes, tout simplement, mais les tableaux apparaissent comme des lignes où les colonnes sont séparées par des tabulations.

On pourrait se contenter de cette mise en page sommaire, si on est très pressé. Cependant, des lignes découpées par des tabulations sont une façon normale de transporter des tableaux. On peut donc sélectionner les lignes qui composent le tableau (pas la précédente ni la suivante) :

(filières·formation)¶

| | effectifs | %/Total |
|---|---|---|
| prépro3R | 113 | 20.62%¶ |
| préproBât | 126 | 22.99%¶ |
| foGénéBase | 197 | 35.95%¶ |
| alpha | 112 | 20.44%¶ |
| Total | 548 | 100.00%¶ |

Efficacité·entropique·:·97.8%¶

puis Utiliser le menu Tableau/Convertir/Texte en tableau/Option ajuster au contenu. On obtient un authentique tableau avec des lignes :

| | effectifs | %/Total |
|---|---|---|
| prépro3R | 113 | 20.62% |
| préproBât | 126 | 22.99% |
| foGénéBase | 197 | 35.95% |
| alpha | 112 | 20.44% |
| Total | 548 | 100.00% |

L'alignement n'est pas totalement satisfaisant : on pourrait y remédier en cadrant tout à gauche, ou autre manœuvre de ce genre. Il faut cependant savoir, pour comprendre ce qu'on fait, que les tableaux d'Hector sont composés dans la police de caractère `Courier New en 10 points`, et que c'est ainsi qu'on les trouve quand on passe par la méthode du Document de transport. Cette police a comme caractéristique principale d'être à *chasse fixe*, ce qui signifie que tous ses caractères ont la même largeur : il est donc facile d'aligner des textes en se servant de cette propriété. Si on applique cette police au tableau, on obtient ceci :

| | effectifs | %/Total |
|---|---|---|
| prépro3R | 113 | 20.62% |
| préproBât | 126 | 22.99% |
| foGénéBase | 197 | 35.95% |
| alpha | 112 | 20.44% |
| Total | 548 | 100.00% |

Ce qui est conforme au résultat vu à l'écran d'Hector. Maintenant, si ça ne plaît pas, on peut faire ce qu'on veut : c'est du vrai tableau éditable et modifiable. On peut par exemple lui appliquer un style de tableau tout fait, avec la commande Tableau/Format automatique :

| | effectifs | %/Total |
|---|---|---|
| prépro3R | 113 | 20.62% |
| préproBât | 126 | 22.99% |
| foGénéBase | 197 | 35.95% |
| alpha | 112 | 20.44% |
| Total | 548 | 100.00% |

C'est entièrement affaire de goût.

Les choses peuvent se corser avec un de ces tableaux qui, dans Hector, nécessitaient deux lignes par case.

L'application de la méthode ci-dessus donne ceci :

| N %L<br>+ | tertiaire | fi courtes | fo gé base | habitat | alpha | S/LIGNE : |
|---|---|---|---|---|---|---|
| néant | 11 14%<br>-- | 6 7%<br>- | 33 41% | 11 14% | 20 25%<br>++ | 81 100% |
| CEB | 3 19% | 4 25% | 6 38% | 2 12% | 1 6% | 16 100% |
| sec inf | 14 44%<br>+++ | 2 6% | 10 31% | 2 6% | 4 12% | 32 100% |
| sec sup | 21 24% | 16 18%<br>+ | 26 30% | 11 13% | 13 15% | 87 100% |
| S/COLONNE: | 49 23% | 28 13% | 75 35% | 26 12% | 38 18% | 216 100% |

Alors que le résultat souhaité serait plutôt cela :

| N %L<br>+ | tertiaire | fi courtes | fo gé base | habitat | alpha | S/LIGNE : |
|---|---|---|---|---|---|---|
| néant | 11 14%<br>-- | 6 7%<br>- | 33 41% | 11 14% | 20 25%<br>++ | 81 100% |
| CEB | 3 19% | 4 25% | 6 38% | 2 12% | 1 6% | 16 100% |
| sec inf | 14 44%<br>+++ | 2 6% | 10 31% | 2 6% | 4 12% | 32 100% |
| sec sup | 21 24% | 16 18%<br>+ | 26 30% | 11 13% | 13 15% | 87 100% |
| S/COLONNE: | 49 23% | 28 13% | 75 35% | 26 12% | 38 18% | 216 100% |

Les instruments d'édition des tableaux permettent d'enlever une à une, mais assez rapidement, les lignes superflues, mais on peut aussi utiliser un format automatique qui colore différemment un couple de lignes sur deux : ici le format Liste2, utilisable à condition d'avoir au préalable supprimé la ligne placée sous la ligne de titre :

| N %L | tertiaire | fi courtes | fo gé base | habitat | alpha | S/LIGNE : |
|---|---|---|---|---|---|---|
| néant | 11 14%<br>-- | 6 7%<br>- | 33 41% | 11 14% | 20 25%<br>++ | 81 100% |
| CEB | 3 19% | 4 25% | 6 38% | 2 12% | 1 6% | 16 100% |
| sec inf | 14 44%<br>+++ | 2 6% | 10 31% | 2 6% | 4 12% | 32 100% |
| sec sup | 21 24% | 16 18%<br>+ | 26 30% | 11 13% | 13 15% | 87 100% |
| S/COLONNE: | 49 23% | 28 13% | 75 35% | 26 12% | 38 18% | 216 100% |

L'idéal serait de pouvoir définir ses propres formats de tableau. Dans la version de Word © dont on dispose ici, ce n'est pas le cas, mais si ça se trouve, ça l'est déjà dans des versions plus récentes.

En tous cas ça l'est dans OpenOffice Writer 3.0 :

| N %L | tertiaire | fi courtes | fo gé base | habitat | alpha | S/LIGNE : |
|---|---|---|---|---|---|---|
| + | | | | | | |
| néant | 11 14% | 6 7% | 33 41% | 11 14% | 20 25% | 81 100,00% |
| | -- | - | | | ++ | |
| CEB | 3 19% | 4 25% | 6 38% | 2 12% | 1 6% | 16 100,00% |
| | | | | | | |
| sec inf | 14 44% | 2 6% | 10 31% | 2 6% | 4 12% | 32 100,00% |
| | +++ | | | | | |
| sec sup | 21 24% | 16 18% | 26 30% | 11 13% | 13 15% | 87 100,00% |
| | | + | | | | |
| S/COLONNE: | 49 23% | 28 13% | 75 35% | 26 12% | 38 18% | 216 100,00% |

Ici (copie d'écran) on a utilisé un format Tableau2 créé une seule fois à la main et réemployé ensuite, où une ligne sur deux n'a pas de trait horizontal, et où la police de caractères est Courier New 10 points. En une seule manœuvre, on parvient à Convertir le tableau en texte et à sélectionner le format Tableau2. Quelques fantaisies apparaissent parfois dans le traitement des espaces : il y en a qui disparaissent, pour remettre un espace dans un tableau il faut en taper deux (*don't ask me why !*), et parfois il faut user de l'outil d'optimisation des colonnes. Dans l'ensemble, c'est assez rapide.

Si on n'a pas très envie de se livrer à ces gymnastiques, il y a aussi la solution de coller dans Excel © :

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | (niveau formation) x (filières FUNOC) | | | | | | |
| 2 | N %L | tertiaire | fi courtes | fo gé base | habitat | alpha | S/LIGNE : |
| 3 | + | | | | | | |
| 4 | néant | 11 14% | 6 7% | 33 41% | 11 14% | 20 25% | 81100% |
| 5 | | -- | - | | | ++ | |
| 6 | CEB | 3 19% | 4 25% | 6 38% | 2 12% | 1 6% | 16100% |
| 7 | | | | | | | |
| 8 | sec inf | 14 44% | 2 6% | 10 31% | 2 6% | 4 12% | 32100% |
| 9 | | +++ | | | | | |
| 10 | sec sup | 21 24% | 16 18% | 26 30% | 11 13% | 13 15% | 87100% |
| 11 | | | + | | | | |
| 12 | S/COLONNE | 49 23% | 28 13% | 75 35% | 26 12% | 38 18% | 216100% |
| 13 | | | | | | | |
| 14 | Avec 6 correction(s) de Yates, Khi2 = 20.93 pour 12 d.d.l. s. à .10 | | | | | | |

Bonne surprise, les tabulations sont détectées, et ça s'organise spontanément en tableau, qu'il reste à copier-coller vers le traitement de texte. On retrouve le petit défaut sur les espaces dans la dernière colonne.

OpenOffice Calc accepte également ce copier-coller direct : il y met un peu plus de formalité avec un écran intermédiaire de confirmation, mais le résultat est le même.

On notera que ce copier-coller direct dans les tableurs permet aussi d'utiliser, si on le désire et si on les trouve belles, d'autres formes graphiques que celles d'Hector.

Cela ne présente pas grand intérêt de coller du texte et des tableaux dans un éditeur graphique ni dans une présentation, aussi n'insistera-t-on pas sur Draw, ni Impress.

*Avec Alt-clic-gauche, copie spéciale*

Dans certaines sections d'Hector, telles que les tris collectifs, mais aussi les pages SPECIALITES et FACTORIELLES, certains tableaux sont des pseudo-tableaux, au sens où ils sont mis en page sans tabulations, mais avec des espaces :

```
Analyse de la collection numérique ou logique {t1G3}
_____________________________________________________________________
     moyenne   écart-type      r(i,T-i)       item
        0,49         0,50       0,613***      ''24 A:faire de nouveaux amis
        0,45         0,50       0,563***      ''27 A:pas tout seul à problèmes
        0,44         0,50       0,595***      ''26 A:oblige à sortir de chez soi
        0,45         0,50       0,569***      ''28 A:parler avec des gens
        0,54         0,50       0,574***      ''30 A:plus de confiance en moi
        0,55         0,50       0,604***      ''35 A : apprendre beaucoup
        0,44         0,50       0,521***      ''9 O : ça va me faire du bien
        0,53         0,50       0,560***      ''10 O:le conjoint est d'accord
        0,44         0,50       0,569***      ''13 O:mon conjoint m'encourage

Corrélations item/test (cohérence) : min, moy, max : 0,521, 0,574, 0,613
Alpha de Cronbach (fiabilité) = 0,854
Alpha comparable pour 10 items = 0,866
```

Copiés collés avec Ctrl-clic-gauche, on obtient ceci (copie d'écran dans Word) :

Analyse·de·la·collection·numérique·ou·logique·{t1G3}¶

_____

·····moyenne···écart-type·····r(i,T-i)·····item¶
·····0,49·····0,50·····0,613***·····"24·A:faire·de·nouveaux·amis¶
·····0,45·····0,50·····0,563***·····"27·A:pas·tout·seul·à·problèmes¶
·····0,44·····0,50·····0,595***·····"26·A:oblige·à·sortir·de·chez·soi¶
·····0,45·····0,50·····0,569***·····"28·A:parler·avec·des·gens¶
·····0,54·····0,50·····0,574***·····"30·A:plus·de·confiance·en·moi¶
·····0,55·····0,50·····0,604***·····"35·A·:·apprendre·beaucoup¶
·····0,44·····0,50·····0,521***·····"9·O·:·ça·va·me·faire·du·bien¶
·····0,53·····0,50·····0,560***·····"10·O:le·conjoint·est·d'accord¶
·····0,44·····0,50·····0,569***·····"13·O:mon·conjoint·m'encourage¶
·¶
Corrélations·item/test·(cohérence)·:·min,·moy,·max·:·0,521,·0,574,·0,613¶
Alpha·de·Cronbach·(fiabilité)·=·0,854¶
Alpha·comparable·pour·10·items·=·0,866¶

Il ne s'agit pas d'un véritable tableau. On peut parvenir à le mettre en forme ainsi, mais ça va être fastidieux.

Avec Alt-clic-gauche, on obtient ceci :

Analyse·de·la·collection·numérique·ou·logique·{t1G3}¶
moyenne → écart-type → r(i,T-i)·item¶
0,49 → 0,50 → ·0,613*** → "24·A:faire·de·nouveaux·amis¶
0,45 → 0,50 → ·0,563*** → "27·A:pas·tout·seul·à·problèmes¶
0,44 → 0,50 → ·0,595*** → "26·A:oblige·à·sortir·de·chez·soi¶
0,45 → 0,50 → ·0,569*** → "28·A:parler·avec·des·gens¶
0,54 → 0,50 → ·0,574*** → "30·A:plus·de·confiance·en·moi¶
0,55 → 0,50 → ·0,604*** → "35·A·:·apprendre·beaucoup¶
0,44 → 0,50 → ·0,521*** → "9·O·:·ça·va·me·faire·du·bien¶
0,53 → 0,50 → ·0,560*** → "10·O:le·conjoint·est·d'accord¶
0,44 → 0,50 → ·0,569*** → "13·O:mon·conjoint·m'encourage¶
Corrélations·item/test·(cohérence)·:·min,·moy,·max·:·0,521,·0,574,·0,613¶
Alpha·de·Cronbach·(fiabilité)·=·0,854¶
Alpha·comparable·pour·10·items·=·0,866¶

Cette fois, la mise en page est faite de tabulations : Hector a remplacé tout groupe de plus de deux espaces par une tabulation, a supprimé les espaces initiaux et les lignes de caractères 'souligné'. On peut donc , en convertissant en tableau, puis en cadrant à droite là où c'est utile, obtenir quelque chose de ce genre :

| moyenne | écart-type | r(i,T-i) | item |
|---|---|---|---|
| 0,49 | 0,50 | 0,613*** | "24 A:faire de nouveaux amis |
| 0,45 | 0,50 | 0,563*** | "27 A:pas tout seul à problèmes |
| 0,44 | 0,50 | 0,595*** | "26 A:oblige à sortir de chez soi |
| 0,45 | 0,50 | 0,569*** | "28 A:parler avec des gens |
| 0,54 | 0,50 | 0,574*** | "30 A:plus de confiance en moi |
| 0,55 | 0,50 | 0,604*** | "35 A : apprendre beaucoup |
| 0,44 | 0,50 | 0,521*** | "9 O : ça va me faire du bien |
| 0,53 | 0,50 | 0,560*** | "10 O:le conjoint est d'accord |
| 0,44 | 0,50 | 0,569*** | "13 O:mon conjoint m'encourage |

Ce qui est déjà beaucoup plus propre. Si on souhaite plutôt conserver la mise en forme d'origine, ne pas oublier de mettre le texte en Courier New, 10 points (ou moins).

Cette variante ne change rien aux « vrais tableaux ».

## Limites et restrictions

Le nouveau copier-coller est susceptible de faire gagner pas mal de temps, notamment dans une démarche où en élabore le commentaire parallèlement à l'investigation statistique. Le fin du fin est de travailler Hector sur un ordinateur portable, auquel on a ajouté un écran auxiliaire sur lequel on met le traitement de texte. Passer du calcul à l'écriture revient alors à passer d'un écran à l'autre.

Cependant, il n'est pas absolument possible de faire tout ce qu'on faisait avec le Document de transport.

Ainsi, si les copier-coller fonctionnent parfaitement pour les pages TRAITEMENTS, FACTORIELLES et SPECIALITES, ils ne fonctionnent pas dans la page SEQUENCES (en tous cas, pas dans cette première version).

De plus, dans la page TRAITEMENTS, certaines fonctionnalités ne sont pas conservées par le copier-coller des tableaux : c'est le cas de la coloration des tableaux de contingence selon le signe des associations locales, ainsi que du découpage des tableaux trop grands. Dans les deux cas, l'utilisateur sera amené, si besoin est, à faire le travail à la main dans le traitement de texte, ou à passer par la procédure du Document de transport.

# TABLE DES MATIERES


Tris, croisements et traitements simples ..... 2
Aspect de la page TRAITEMENTS ..... 2
Parties communes ..... 2
Description du panneau de commande [tris] ..... 3
Combinaisons de types dans les traitements simples ..... 5
Tris d'une seule variable ..... 5
Tris de deux variables ..... 13
Tris de plus de deux variables ..... 21
Les filtres ..... 23
Exporter les résultats ..... 25

Traitements collectifs ..... 26
La page TRAITEMENTS, volet [collectifs] ..... 26
Tris en série ..... 27
Matrices de statistiques ..... 28
Analyse de tests ..... 31
Discrimination et difficulté ..... 31
Cohérence, fiabilité ..... 34

Le plan de projection ..... 36
La page TRAITEMENTS, volet [projection] ..... 36
Les axes du plan et ce qu'on y projette ..... 38
Projection des sujets ..... 38
Projection des variables ..... 40
Mise en œuvre ..... 41

Exploitation des résultats ..... 43
La notion de document ..... 43
Document au format .rtf ..... 44
Sorties sans mise en forme ..... 45
Sortie avec mise en forme ..... 46
Document au format .odt ..... 47
Document au format .tex ..... 47
En-tête du document ..... 48
Statistiques globales ..... 49
Tableaux ..... 49
Graphiques ..... 49
Document au format .html ..... 50

L'impression directe ..... 50
Le copier-coller ..... 50
Les variantes du copier-coller ..... 51
Copier-coller une image seule ..... 51
Coller textes et tableaux ..... 55
Limites et restrictions ..... 60

Table des matières ..... 61